Toward Evaluation that Leads to Best Practices: Reconciling Dialog Evaluation in Research and Industry

Tim Paek

Microsoft Research
One Microsoft Way
Redmond, WA 98052
timpaek@microsoft.com

Abstract

Dialog evaluation is approached in different ways by research and industry. While researchers have sought commensurable evaluation metrics that allow for comparison of disparate systems with varying tasks and domains, industry engineers have focused mostly on best practices and delivering a return-on-investment to customers. In this paper, we contend that the problem of finding commensurable metrics also applies to commercial evaluation, and critically survey four candidate metrics for commensurability. Finally, in light of the problems faced by the candidate metrics, we advocate a collaborative agenda for dialog evaluation based on using statistical meta-analysis for empirically establishing best practices from any evaluation metric.

1 Introduction

Since the beginning of speech recognition research, which started more than 50 years ago, people have dreamed of being able to talk to machines and appliances as if they were human. What began in academic institutions and industry laboratories gradually made its way into the marketplace around the mid 1990s with the commercial introduction of voice user interfaces (VUI) (Pieraccini & Lubensky, 2005). Most VUI applications were spoken dialogue systems for automating customer service tasks. Nowadays, hundreds of commercial systems are being deployed by companies each year, adhering to industry-wide standards and protocols established to ensure the interoperability of components and vendors, such as VoiceXML, CCXML, MRCP, etc. Unfortunately, as some researchers have pointed out (Pieraccini & Huerta, 2005), dialog systems in industry have been evolving on a parallel path with those in academic research, where usability and cost have been the primary goal of industry, and naturalness of interaction and freedom of expression have been the goal of research. Given that commercial systems are now beginning to embrace less-constrained interaction models, a call has been made for a “synergistic convergence” of architectures, abstractions and methods from both communities, lest research results become irrelevant to industry practice (Pieraccini & Huerta, 2005).

It is under this motivation that we critically survey the field of dialog evaluation in research and industry[1]. Dialog evaluation is a task common to both communities, but it has been approached in distinct ways. Research has exerted considerably effort and attention to devising evaluation metrics that allow for comparison of disparate systems with varying tasks and domains. In industry, system engineers generally do not gruel over what metric is best-suited for comparison of disparate systems. Because companies live and die by practical evaluation, what matters most is that they improve their customers’ business; hence, beyond measures that evince return-on-investment (ROI), there is little focus on dialog evaluation metrics.

This paper endeavors to bridge the gap in understanding of dialog evaluation between academic research and industry. We explore what research and industry has to learn from each other, and how working together can advance the goals of both communities. We do this in three sections. In the first section, we describe differences in the way

[1] Though lines are often blurred, by “research” we mean academic institutions and industry laboratories whose focus is not immediate commercial gain.

Bridging the Gap: Academic and Industrial Research in Dialog Technologies Workshop Proceedings, pages 40–47,
NAACL-HLT, Rochester, NY, April 2007.

evaluation is pursued by both communities. In particular, we discuss the academic search for commensurable metrics that allow for comparison of disparate systems. In considering evaluation in industry, we expound on the drive for VUI best practices. In the second section, we survey four candidate metrics for commensurability. Irrespective of whether they achieve that purpose, they are likely to be of practical interest to system engineers in industry. Finally, in the last section, we propose a collaborative agenda for dialog evaluation to advance the goals of both communities. In particular, we advocate statistical meta-analysis for empirically establishing best practices from any dialog evaluation metric.

2 Differences in Approach

In research, dialog evaluation is considered a hard problem. Several workshops and special issues of journals have already been devoted to this topic. On the other hand, in industry, dialog evaluation is considered relatively straightforward. Any dialog system worth the effort of deployment must ultimately deliver ROI. That ROI may be in terms of the cost savings accrued from automating what is typically handled by human operators, such as in call centers, or in terms of expanding the breadth and depth of customer service that has typically been privy to large enterprises. The focus in industry has not been so much on how best to evaluate a dialog system, but rather on how best to design them. In short, industry tends to focus more on best practices than on evaluation.

The difference in thinking cannot be fully attributed to the difference in goals propounded earlier (Pieraccini & Huerta, 2005); namely, that industry generally pursues usability and cost-effectiveness, whereas research pursues unconstrained spoken interaction under the assumption that usability would naturally follow. There is also a lack of understanding about what dialog evaluation in research has to offer industry, and in particular, for the kind of directed dialogs (which restrict what users can say but generally improve usability) that are common in commercial systems. In this section, we explore why dialog evaluation is considered so challenging in the research community, and demonstrate how many of the same issues that researchers face are also applicable to VUI engineers as well.

2.1 Commensurability Problem

Observing the success that both the speech recognition and spoken language understanding communities have enjoyed in advancing their technologies by establishing a controlled, objective and common evaluation framework (Pieraccini & Lubensky, 2005), dialog researchers have sought a similar evaluation framework for their work. This framework could not only be used to gauge technological progress in the field but also allow for the assessment of diverse systems of varying tasks and domains. Unfortunately, the research community has yet to agree upon such a framework. To date, researchers operate under on a variety of different frameworks, and new evaluation metrics are proposed all the time.

Part of the reason for this has to do with the complexity of the evaluation task. On the one hand, dialog systems are ultimately created for users, so usability factors such as satisfaction or likelihood of future use should be primary. On the other hand, because usability factors are subjective, they can be erratic and highly dependent on the complex interplay of user interface attributes (Kamm et al., 1999). So, designers have turned to objective metrics such as task completion time or dialog success rate (e.g., see Gibbon et al., 1998 for review). Due to the interactive nature of conversation however, these metrics do not always correspond to the most effective user experience (Hartikainen et al., 2004; Lamel et al., 2000). Objective evaluation of user experience can itself be highly uncertain or non-existent (Dybkjaer & Bernsen, 2001). Furthermore, in many cases, it is just not clear how to apply an objective metric. Even an ostensibly straightforward metric, such as task success, can be difficult to ascertain. For example, defining the “success” of a session with an intelligent tutoring system is no easy task, and may or may not have anything to do with student learning, depending on what constitutes the basis for comparison (either a human or a keyboard system).

The choice of evaluation metric depends on the purpose of the evaluation (Dybkjaer & Bernsen, 2001; Paek, 2001). Some researchers are more interested in achieving human-human conversation-like qualities in their systems than others. Because researchers have different purposes, they have developed a wide assortment of dialog evaluation metrics. As mentioned earlier, metrics can be subjective or objective, deriving from questionnaires or log files. They can vary in scale from the utterance level to the overall dialog (Glass et al., 2000). They can treat the system as a "black box" and describe only its external behavior (Eckert at al., 1998), or as a "glass box" and detail its internal processing. If one metric fails to suffice, several metrics can be combined (Walker et al., 1997). Finally, if all else fails to suffice, then new metrics can be developed.

Despite the diversity of metrics and purposes, researchers have wanted to compare their dialog systems against others. They have sought an evaluation metric or framework that could facilitate comparative judgments. In philosophical terms, they are seeking a measure of commensurability; two quantities are commensurable if both can be measured by the same units. But what units can allow one dialog system to be compared against another when they vary along so many different dimensions, such as components, interface attributes, domains and tasks? These different aspects can also interact with each other in highly complex ways.

Commensurability is not only a problem for research, but also for industry as well. Suppose that a commercially deployed system adhering to VUI best practices allows a customer to achieve a 90% task completion rate and a savings of $500 million dollars. Because the system consists of many architectural and interface attributes that may interact with one another, from the exact wording of the prompts to the dialog management strategies employed, how can system engineers really know if they found the optimal configuration? Perhaps given a new set of dialog strategies, or slightly different prompt wording, task completion could be significantly improved. The issue of commensurability still applies because ideally engineers would like to be able to say that the system they built, with the configuration that they arrived at, is somehow better than other systems that they, or even their competitors, could have designed. Of course, free market economics could be the judge, and engineers who provide higher ROI might be able to stake their claim on superiority. However, the factors that play a role in making a dialog system usable and efficacious can be multifaceted, and may even reach beyond the choices of the designer to the characteristics of the user population, or user profile, and usage patterns (Frostad, 2003).

2.2 Best Practices

System engineers in industry have implicitly dealt with the issue of commensurability by relying on VUI best practices. Best practices often emerge through trial-and-error and the test of time, and essentially serve as de facto industry standards (Balentine & Morgan, 2001). The need for best practices evolved from classical software engineering principles. Because system engineers were responsible for the complete specification of system behavior, they began applying software engineering principles such as requirements gathering, specification, design and coding, usability testing, and post-deployment tuning to make sure that their systems could scale into high-quality, commercial grade solutions (Pieraccini & Huerta, 2005). Along the way, engineers encountered problems, and as they began to notice the same problems appearing over again, they began to devise best practices for the design and deployment of VUI systems. As the industry has matured over the years, best practices have been collected into books (e.g., Balentine & Morgan, 2001; Cohen et al., 2005), and many platform providers offer seminars, training, and online resources for learning best practices (e.g., Frostad, 2003).

It is important to note that best practices are not the sole propriety of industry alone. Academic researchers, who have been building a multitude of systems under various government sponsored projects, such as DISC and DISC-2, have also developed their own best practices (e.g., Lamel et al., 2000; Dybkjaer & Bernsen, 2001).

Best practices often come in the form of practical dos and don'ts. For example, for telephony-based systems, almost all published literature in industry and research recommends that prompts be kept short and simple. Sometimes these practices, such as this one for prompts, are validated either directly or indirectly by experimental design. Various academic institutions have also pursued VUI design experiments, and published their findings, which often get cited in industry. For example, both the Dialogue Engineering Project[2] at CCRI, University of Edinburgh and the Stanford CHIMe Lab[3] are well-known to industry (e.g., Nass & Brave, 2005).

[2] http://www.ccir.ed.ac.uk/doc/ccir_dialogues.htm

[3] http://chime.stanford.edu/

The problem with best practices is that they are often not substantiated by rigorous experimental design, which is why the previously mentioned academic institutions have sought to conduct their research. In the worst case, best practices are based on the accumulated experience and intuition of system engineers and consultants, which unfortunately is prone to error and cannot be generalized beyond their limited experience. This can result in ostensibly contradictory recommendations. For example, whereas one best practice may advocate personifying the dialog system using the first person singular, another may advocate adhering as close as possible to a non-personified, touch-tone model. In this particular case, the contradiction stems from limited knowledge of the technologies available at the time; the first was made with HMIHY technology (reference) for mixed-initiative interaction in mind, whereas the other was not.

Even when best practices are based on experimental studies, they cannot be automatically generalized beyond the conditions and assumptions of the experimental design. Controlled experimental design dictates that in order to find a significant effect of a treatment, such as prompt wording or gender of voice, other factors should be held constant, such as the dialog flow of the system. When those other factors change, the effect of the treatment may change as well. Hence, results cannot be automatically generalized as best practices beyond their experimental settings. Furthermore, as many of the studies themselves point out, they are limited to the characteristics of their subject population. In fact, a common industry best practice is to conduct pilot usability studies on the domain task to better understand the needs and usage patterns of the expected user population (Balentine & Morgan, 2001).

The point here is not to discourage the use of best practices, but to highlight the need for rigorous validation of them. Incommensurability poses a problem for best practices because when disparate systems cannot be evaluated according to a common framework, it is hard to generalize system features or attributes into best practices. Ideally, dialog evaluation should foster the development of best practices.

3 Survey of Commensurable Metrics

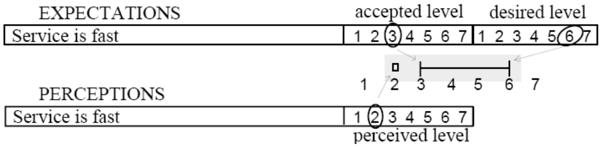

Figure 1. A graphical display of the gap between user expectations and perceptions for the SERVQUAL method.

In the research literature, several dialog evaluation metrics have been proposed which in some fashion or another could allow for the comparison of disparate systems. In this section, we critically survey four such metrics. Although other metrics might qualify as well, these metrics were chosen because they are likely to be of practical interest to system engineers in industry, regardless of whether they facilitate commensurability.

3.1 SERVQUAL

SERVQUAL is a SERVice QUALity evaluation method developed by marketing academics and applied to spoken dialogue systems by Hartikainen et al. (2004). SERVQUAL consists of a questionnaire and methods of data analysis. The questionnaire[4] provides a subjective measure of the gap between expectations and perceptions in five service quality dimensions: tangibles, reliability, responsiveness, assurance and empathy. Once questionnaire data is collected, two measures, a Measure of Service Superiority (MSS = Perceived level – Desired Level) and a Measure of Service Adequacy (MSA = Perceived Level – Acceptable Level), can be easily computed. Figure 1 shows how these two measures can then be used to display a "zone of tolerance" for users (Hartikainen et al., 2004). Graphical plots showing the relationship between performance and importance are also commonly used.

The SERQUAL method could be considered a commensurable metric because it evaluates the usability of dialog systems with respect to a common unit of measurement; namely, the gap between user expectations and perceptions. Even if disparate systems engender different expectations in users, perhaps because of dissimilar tasks and domains, they can still be compared against each with respect to how far off the reality of their perceived performance is from user expectation.

[4] Accessible online at http://www.cs.uta.fi/hci/spi/SERQUAL/.

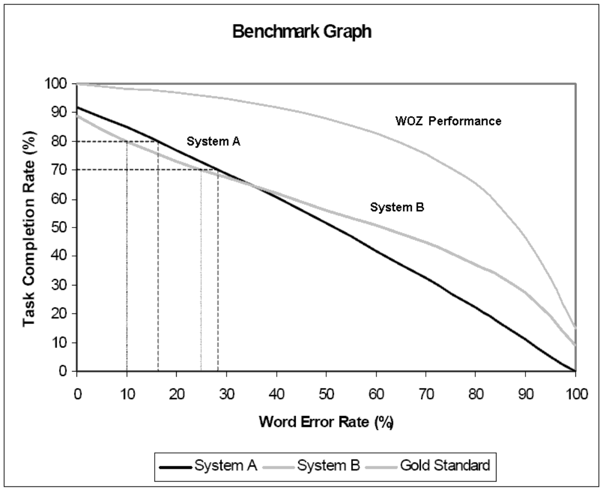

Figure 2. Comparison of task completion rate between two systems and a WOZ gold standard.

The primary problem with SERQUAL is that user perceptions and expectations can be unstable and easily susceptible to manipulation. For example, in conducting numerous experiments on user perception of speech interfaces, Nass and Brave (2005) note: "Reminding people that they depend on the interface for their success automatically makes the computer seem more intelligent… labeling a part of an interface as a specialist, conforming to gender stereotypes, flattering the user, or matching the user's personality also increases perceived competence. Indeed, people are so susceptible to manipulation that perceived intelligence is a very weak predictor of actual intelligence" (p.152). Without fully understanding the subtle factors that can easily influence user perception, it is possible to attribute service quality superiority to the wrong factor. A final concern is that focusing on user perceptions in evaluation may detract some from working out more serious technical flaws.

Despite the problems with SERQUAL, it has a strong tradition in marketing and may appeal to those in industry who need to pitch the value of their systems from a customer service standpoint. Researchers and engineers alike can also benefit from using SERVQUAL because it draws attention to how user perceptions can thwart even well-designed systems.

3.2 WOZ Gold Standard

Whereas SERQUAL measures the gap between user expectations and perceptions, Paek (2001) advocates using human performance in Wizard-of-Oz (WOZ) experiments as a gold standard for benchmarking systems. The idea is that once an evaluation metric (e.g., task completion) is selected, a WOZ experiment can be conducted that compares different treatments of interest (e.g., dialog repair strategies) along varying levels of word-error rate. The human wizards in these experiments never hear users directly, but instead receive output mediated by the recognizer and/or spoken language understanding unit. Although other researchers have recommended similar WOZ setups (Stuttle et al., 2004), the focus here is on quantifying the difference in performance between the system and a human gold standard, and using that as a commensurable metric. For example, Figure 2 displays the task completion rates of two dialog systems as well as a human wizard. The performance of each system is measured as the difference in density between that system and the WOZ. Notice that depending on the interval of interest in word-error rate, system A can be better than B and vice versa. The difference in density between the WOZ performance and the absolute upper-bound represents the difficulty of the domain task for human wizards, or the "benchmark complexity" (Paek, 2001).

Using human wizards as a gold standard allows for the comparison of disparate systems. If the difference between the performance of any system and a human wizard is small, then that might suggest that the system is performing very well for that domain task, regardless of what that domain task is itself. However, if the benchmark complexity is also small, that would suggest that the system may be performing well because the task is easy.

The major problem with using human wizards as a gold standard is the effort required to conduct WOZ experiments. It is not only time-consuming and costly, but technically challenging to insert a wizard into the right place in the processing of utterances and to make sure that they can effectively do their job. Once a wizard is in place, it can also be difficult to obtain data points along a wide range of word-error rates.

Despite these problems, having a human gold standard naturally lends itself to optimization, which is always of interest to industry. System engineers can identify which components are contributing the most to a performance metric by examining the density differences with and without particular components. Furthermore, if customers are willing to identify how much they might be willing to pay to achieve various levels of performance – i.e., if their utility functions are elicited, then it is possible to calculate average marginal costs by weighting density differences by their corresponding utilities (Paek, 2001).

3.3 SASSI

Noting the lack of psychometric validation of subjective usability measures often used in evaluations of spoken dialog systems, Hone & Graham (2000) propose a questionnaire measure for the Subjective Assessment of Speech System Interfaces (SASSI). They identify four weaknesses of subjective evaluation measures (e.g., questionnaire items), which are worth repeating here. First, the content and structure of these measures are for the most part arbitrary and based on intuition. Second, measures are not validated against other subjective or objective measures. This renders their construct validity suspect; that is, it difficult to tell if they really measure what they are intended to measure. Third, these measures do not report their reliability, both in terms of their test-retest stability across time, as well as the internal consistency of a group of measures for a particular construct. Finally, measures are commonly summed or averaged to obtain an overall score when such an approach can only be justified on the basis of evidence that all of the measures really do assess the same construct (Hone & Graham, 2000).

In order to build a psychometrically valid, reliable and sensitive questionnaire, SASSI was developed by first taking questionnaire items from established measures in the research literature, and then having users respond to those items with respect to four different speech applications. After collecting the data, exploratory factor analysis was conducted to find a set of theoretical constructs, or "factors". These factors are represented by a set of questionnaire items which tend to be highly correlated with each other. Six main factors in users' perceptions of speech applications were identified:

- System Response Accuracy: User's perceptions of the system as accurate and doing what they expect.
- Likeability: User's rating of the system as useful, pleasant and friendly.
- Cognitive Demand: The perceived amount of effort needed to interact with the system and feelings arising from this effort.
- Annoyance: User's rating of the system as repetitive, boring, irritating and frustrating.
- Habitability: The extent to which users knew what to do and what the system was doing.
- Speed: How quickly the system responded to user inputs.

It is important to note that this factor analysis was preliminary and did not benefit from multiple iterations. Hence, only the first three factors had internal consistency reliabilities, as measured by Cronbach's alpha, of $\alpha \geq 0.80$, which is typically required for widespread adoption.

Although the development of SASSI is definitely a promising start for creating valid and reliable subjective measures, it does not say much about how system features, such as prompt wording, influence the six factors identified. In other words, SASSI only tells system engineers what to measure (which is an important contribution), not how to design their systems. For that, statistical analyses relating system features to SASSI measures are needed. Nevertheless, the SASSI methodology represents a valuable, principled way of determining common units of measurement for comparing disparate systems. We return to this issue again in Section 4.

3.4 PARADISE

Perhaps the best-known general framework in research for dialog evaluation is PARADISE (PARAdigm for Dialogue System Evaluation) (Walker et al., 1997). PARADISE addresses three goals: 1) to support the comparison of multiple systems on the same domain task, 2) to provide a method for developing predictive models of user satisfaction as a function of system features, and 3) to provide a technique for making generalizations across systems about which features impact usability (Walker et al., 2000). Treating user satisfaction as the primary objective function, PARADISE derives a combined performance metric as a weighted linear combination of task-success measures and dialog costs, the latter consisting of two types: dialog efficiency metrics (e.g., elapsed time), and dialog quality metrics (e.g., mean recognition score). Deriving the metric simply involves model-building using multivariate linear regression.

Although PARADISE is geared towards comparing systems that perform the same domain task, it does provide a general framework for at least comparing those systems. The problem is that while PARADISE is a useful descriptive tool, its power to generalize has been somewhat limited. In pursuing the third goal of generalization – to figure out what system features really matters to users, PARADISE was applied to experimental data from three different dialog systems (Walker et al, 2000). Models trained on one system were then tested on the other two systems. Results showed that the models do indeed generalize well across the three systems. However, the three features that consistently appeared among the top predictive factors were mean recognition score, whether users reported that they had completed the task, and the percentage of recognition rejections. Unfortunately, this is not the kind of insight that leads to best practices, and most system engineers probably already knew that improving speech recognition and task completion (either in absolute terms or by user perception) would improve user satisfaction.

What is likely to be of practical interest to system engineers in industry about PARADISE is the usefulness of performing multivariate linear regression to predict measures of interest based on not only task success but measures of dialog efficiency and dialog quality. Because most of the PARADISE features can be automatically generated from data, apart from having users fill out a satisfaction survey, it is of almost no cost to perform a PARADISE analysis.

4 Evaluation That Leads to Best Practices

In the previous section, we critically surveyed four dialog evaluation metrics that could be considered candidates for commensurability. In light of the problems faced by these metrics, in this Section, we propose a collaborative agenda for dialog evaluation that fulfills the need in industry for best practices and the research pursuit of generalizations. Before considering the proposal, however, we reassess the value of commensurability.

4.1 Reassessing Commensurability

Although commensurability seems to be worthwhile, in looking closely at the desire to compare disparate systems of varying domains and tasks, it is important to separate the question of, "How is the dialog system doing relative to other systems?" from "How can the dialog system do better?" Answering the former question is truly a challenging task, and metrics like SERVQUAL and the WOZ gold standard offer interesting solutions. However, there is little to gain from answering this question other than bragging rights. The latter question of how to improve a dialog system is ultimately more beneficial to research, and does not necessarily require finding a commensurable metric.

In Section 2.1, we argued that commensurability was a problem for industry because system engineers would like to be able to say that the system they built is somehow better than other systems that they, or even their competitors, could have designed. However, system engineers can still say this without having to answer the question of how their system is doing relative to others. If they have established best practices for improving any dimension of their system, they can be assured that they have sought the optimal design.

The claim here is that the research community can benefit from focusing less on the question of relative performance and more on the question of how to improve dialog systems. Instead of trying to find commensurable metrics, we propose that the field should seek empirical and experimental evidence of factors that can improve any dialog metric, such as SASSI, regardless of domain or task. By doing so, the research community has more chance to influence industry best practices.

4.2 Proposal

In order to answer the question of how best to improve dialog systems, we propose pooling data from both research and industry to conduct meta-analyses. Meta-analysis, which is widely used in biomedicine and behavioral sciences, is the statistical analysis of a large collection of results from individual studies for the purpose of integrating the findings (Glass et al., 1981). By synthesizing results of related studies, the combined weight of evidence can be applied.

Meta-analysis for improving dialog systems involves three tasks: attribute identification, data coding, and statistical analysis. Attribute identification entails identifying all attributes of dialog systems that may have any effect on an evaluation metric of interest. For example, minute details such as the gender of the voice output, average and median word length of prompts, average latency to respond, etc. may influence metrics like task completion time. Once attributes have been identified, data pooled from research and industry can be coded by them, and once the data has been coded, it will not only be possible to conduct the kind of psychometric validation and reliability testing of metrics that distinguished SASSI, but also determine through correlation, regression and hypothesis-testing what system attributes influence any particular metric of interest, regardless of domain or task. For example, suppose SASSI scores are collected for a system. For each user interaction, a data entry would consist of the SASSI score, any other evaluation metrics of interest (e.g., task completion time), attributes of the system (e.g., gender of voice) and system interaction (e.g., number of confirmations used), and perhaps even attributes of the user (e.g., age group). Now imagine that every dialog system deployed provides this kind of data. With this data, it would be possible to learn, for instance, that prompts that flatter the user are consistently correlated with high SASSI likeability scores across all commercial and research systems. This provides a basis for empirically establishing best practices.

In order for this proposal to work, a large amount of data is required. Because the number of dialog systems built in research pales in comparison to the hundreds of systems that are commercially deployed in industry each year, researchers must work with system engineers to utilize the same metrics (e.g., the same questionnaires) and to code and pool data. While this task may seem Herculean, the result is of equal benefit to both research and industry: best practices for improving dialog systems that are empirically established.

5 Conclusion & Future Work

In this paper, we have examined the different ways in dialog evaluation is approached in research and industry. We critically surveyed four dialog evaluation metrics that could be considered candidates for commensurability. In light of problems faced by these metrics, and in reassessing the value of commensurability, we proposed a collaborative agenda for dialog evaluation based on using statistical meta-analysis for empirically establishing best practices from any evaluation metric. A meta-analysis is forthcoming as future work.

References

Balentine, B. & Morgan, D. 2001. How to Build a Speech Recognition Application: Second Edition: A Style Guide for Telephony Dialogues, Enterprise Integration Group.

Cohen, M., Giangola, J., Balogh, J. 2004. Voice Use Interface Design. Addison-Wesley Professional.

Dybkjaer, L. & Bernsen, N. 2001. Usability evaluation in spoken language dialogue systems. In ACL Workshop on Evaluation Methodologies for Language and Dialogue Systems.

Eckert, W., Levin, E., & Pieraccini, R. 1998. Automatic evaluation of spoken dialogue systems. In TWLT13, pp.99-110.

Frostad, K. 2003. Best practices in designing speech user interfaces.http://msdn.microsoft.com/library/en-us/dnnetspeech/html/vuibstprcf.asp?frame=true

Gibbon, D., Moore, R. & Winski, R. (Eds.). 1998. Handbook of standards and resources for spoken language systems. Walter de Bruyter, Berlin.

Glass, J., Polifroni, J., Seneff, S. & Zue, V. 2000. Data collection and performance evaluation of spoken dialogue systems: The MIT experience. In Proc. ICSLP.

Glass, G.; McGaw, B.; & Smith, M. 1981. Meta-analysis in Social Research. Beverly Hills: SAGE.

Hartikainen, M., Salonen, E. & Turunen, M. 2004. Subjective Evaluation of Spoken Dialogue Systems Using SERVQUAL Method. In Proc. Interspeech, pp.2273-2276.

Hone, K. & Graham, R. 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). NLE, 6(3-4): 287-303.

Lamel, L., Minker, W., & Paroubek, P. 2000. Towards best practices in the development and evaluation of speech recognition components of a spoken language dialog system. NLE, 6(3-4): 305-322.

Kamm, C. Walker, M. & Litman, D. 1999. Evaluating spoken language systems. In Proc. AVIOS.

Nass, C. & Brave, S. 2005. Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship. Cambridge, MA: MIT Press.

Paek, T. 2001. Empirical methods for evaluating dialog systems. In ACL Workshop on Evaluation Methodologies for Language and Dialogue Systems.

Pieraccini, R. & Huerta, J. 2005. Where do we go from here? Research and commercial spoken dialog systems. In Proc. SIGDIAL, pp. 1-10.

Pieraccini, R. & Lubensky, D. 2005. Spoken language communication with machines: the long and winding road from research to business. In Proc. IEA/AIE, pp. 6-15.

Stuttle, M., Williams, J. & Young, S. 2004. A framework for dialogue data collection with a simulated ASR channel. In Proc. Interspeech, pp. 241-244.

Walker, M., Litman, D., Kamm, C., & Abella, A. 1997 PARADISE: A general framework for evaluating spoken dialogue agents. In Proc. ACL/EACL, pp. 271-280.

Walker, M., Kamm, C. & Litman, D. 2000. Towards developing general models of usability with PARADISE. NLE, 6(3-4): 363-377.