Visualizing Multimodal Interactions: Design and Evaluation of Experience Sharing

Peter Pal Boda
Aalto University
Department of Signal Processing and Acoustics
Otakaari 5 A, FI-02150 Espoo, Finland
peter.pal.boda@gmail.com

Akos Vetek
Nokia Labs
FI-02150 Espoo, Finland
akos.vetek@nokia.com

Abstract

Development of multimodal applications is an iterative, complex, and often a rather heuristic process. This is because in multimodal systems the number of interplaying components can be far greater than in an unimodal Spoken Dialogue System. From the developer's perspective, a multimodal system presents challenges and technical difficulties on many levels. In this paper we will describe our approach to one specific component of multimodal systems, the Multimodal Integrator. On the other hand, from the designer's perspective, all components must be fine-tuned to a level that their combined overall performance can deliver the desired experience to end users. In both cases, evaluation and analysis of the current implementation is paramount. Hence, looking into the details while getting a good understanding of the overall performance of a multimodal system is the other key topic.

1 Introduction

Humans unconsciously rely on a set of input and output modalities when communicating, in a seamless and effortless manner, to deliver messages and to exchange information effectively. Multimodal Interaction is an inherently natural form of human-to-human communication.

When the development of multimodal systems is considered, the input and output modality engines are typically readily available with off-the-shelf modules. Their tuning for a specific application is relatively straightforward. Thus a major focus must be on the Multimodal Dialogue Manager (MDM) and on the Multimodal Integration (MMI) modules. Dialogue management is not in the scope of this article, but it is suffice to mention that with the emergence of standardized markup language technologies like EMMA, the Extensible MultiModal Annotation markup language (Johnston et al. 2009), the interaction modeling and development efforts can be significantly formalized.

Multimodal Integration, on the other hand, presents an interesting opportunity for research. Solutions are typically ad-hoc, case-by-case implementations, although a few larger trends can be identified. With the availability of EMMA, even a new evaluation framework was proposed to assess the performance of multimodal fusion engines (Dumas, Ingold, and Lalanne 2009). However, no winning approach has yet emerged.

Evaluation of multimodal systems is also an evolving field. Multimodal interaction represents a complex phenomenon in many different aspects, but numbers alone do not tell the whole truth. Our assumption is that there is room to look, literally, beyond the numbers to capture the bigger picture, but also to enable designers and developers to zoom into the details of what is happening within a state, within turn-taking with multiple modalities, and what wording, gestures or other modality-specific inputs are used by users at a given stage of a dialogue, how the length of a dialogue correlates to the outcome, and so on.

The paper first discusses the basic intention behind Multimodal Integration along with an evaluation alternative. This is followed by a description of a visualization approach for multimodal interaction to support the overall development process. Finally, the paper closes with a conclusion and a look at future work items.

2 Multimodal Integration

The primary goal of the Multimodal Integration component is to combine the contents of multiple incoming channels into one single semantic representation. This single representation is meant to be used by the subsequent MDM to control the flow of the interaction, to execute selected actions or fetch data if necessary, to present it to the user, etc.

MMI can be classified at least along two dimensions, depending on the internal data representation: the difficulty of the task to solve, and the placement of the integration in terms of the processing flow. *Early* and *late fusion* mechanisms refer to the placement of the integration in the data flow. When the type of the implementation considered, *rule-based*, *statistical* and *algorithmic solutions* are typical, or a combination of them, all operating on the symbolic level in the processing flow. In these implementations the incoming modalities are handled by their dedicated recognition engines, and the integration task is carried out only after the individual recognition results are available. A good overview and classification of the known technologies was prepared by Lalanne et al. (2009).

Early integration refers to the combination of highly correlated low-level signals, e.g. audio-visual features (derived from the incoming speech signal and visual features from a lip reading component) where the integration is carried out on the raw feature level before any classification is done.

On the symbolic level, the integration process takes the outcomes of modality-specific recognition engines and aims to combine them into one single semantic representation. The main tasks at this stage are: 1) combine complementary bits of information, 2) resolve redundancy, if there is any, 3) deal with timing issues, and 4) handle ambiguity. There is a wide range of approaches for symbolic level integration, including rule-based implementations (Adler and Davis 2007), unification grammar-based approach (Johnston 1998), agent-based technology (Flippo, Krebs, and Marsic 2003), finite-state transducers (Johnston and Bangalore 2000) and statistical methods e.g. based on voting mechanism by (Wu, Oviatt, and Cohen 1999). The modified dynamic time warping algorithm in a multidimensional space for bimodal inputs combines the advantages of early and late semantic level fusions (Wollmer et al. 2009).

Another statistical approach, relying on Maximum Entropy (ME) based classification, was introduced by Boda (2004). The method is motivated by language understanding, where the incoming gestures represent the context for the verbal inputs. Integration accuracy with ME classification yielded 76.5% and 87% for the 1st and top 3-best recognition candidates, respectively, while with additional contextual information about the gesture type and about the object the gesture pointed at the 1st and top 5-best recognition results were 91.5% and 96.8%, respectively (Boda 2006).

Multimodal Integration transforms the complexity of a multimodal space into an unimodal case. From an evaluation perspective, MMI is considered a classification problem and the typical measures of accuracy or error rates are utilized. However, these figures describe the performance on the overall level and not specifically in a given stage of the dialogue. Imagine, that at some point during the interaction the incoming gesture and speech inputs are integrated incorrectly, resulting in an erroneous interpretation of the user's intention at that turn. If averaged, using overall evaluation metrics, they will never point to the specific issue at a given dialogue turn. The only way to figure out what errors occur in the integration is to go into the log files and to identify the problems with certain gesture and speech input combinations. In other words, one needs to look into the details.

This is exactly the approach proposed next, when the entire multimodal dialogue is visualized in an interactive way.

3 Multimodal Interaction Visualization

Evaluation of interactive systems is a crucial part of the development process. The more complex the implementation, the earlier in the development one needs to get information about the performance of the system and its components. Spoken Dialogue Systems (SDS) and Multimodal Dialogue Systems (MDS) are prime examples of complex implementations. While there have been plenty of work and projects on the evaluation of SDS, e.g. (Walker et al. 1998), and also within EU and DARPA projects like SUNDIAL and ATIS, respectively, there has been less work in the evaluation of multimodal implementations as overall systems.

For SDS, in their seminal paper Simpson and Fraser (1993) argued that it was too simplistic to suppose that a single metric could be used to evaluate a dialogue system, and that it seemed more promising to assume that a dialogue system could be characterized by a set of quantitative results coupled with some qualitative judgements, such as usability and pleasantness. They divided evaluation metrics in to two classes: black box and glass box metrics. *Black box metrics* are concerned with the performance of the entire system. *Glass box metrics*, on the other hand, look inside the system and monitor the performance of the individual components. Thus, glass box metrics are useful diagnostics during system development, while black box metrics are more suited to characterizing the "goodness" of the system at achieving its ultimate objectives (Fraser 1993).

Following this paradigm, the glass box evaluation suggests a view to the parts, although, speaking of only numerically. On the other hand, why not look, literally, behind these numbers, deeper into the details, and see them as part of the bigger picture? What if a visualization of the ongoing interaction were available, down to the performance level of the recognition engines vs. the actual user inputs?

Trends in visualization Recent years have seen the emergence of a new science that aims to handle data in a more democratized way: easier access to data, seamless interpretation, meaningful and aesthetically pleasing visualizations. Specifically, web-based interactive visualization frameworks have matured. Data visualization is about the better understanding of the world and ourselves to provide useful insights and tools to improve our decisions (Yau 2011). The availability of tools to visualize data grew exponentially in the past few years, like R, Processing, D3.js, raphaël.js, paper.js, OpenFrameworks, just to name a few[1].

Earlier work Only few relevant works were found in the literature for interaction visualization. Two of them are mentioned here. Hakulinen, Turunen, and Salonen (2005) introduced a visually motivated development framework, where visual feedback on turn-taking, system responses and the internal dialogue status is provided. Essentially, an extra visual output modality is added to the SDS implementation. Both users and developers can benefit from the distributed, agent-based visualization framework.

Tat and Carpendale (2002) experimented with different techniques and used Bubba Talk to visualize text-based human-human dialogues. The intention was "to give an impression of the spirit and timbre of the conversation." Animation and dynamic techniques were used to represent connections, while color coding for "some aspects of the mood of each speaker."

This work is based on an earlier attempt of visualization of spoken dialogues (Boda 2000), which provided a view to the flow of the interaction within a speaker-independent name dialing application. The implementation enabled developers to explore different dialogue paths in a statistical sense, from state to state based on statistics collected from the recorded logs. Furthermore, an evaluation measure was introduced that combines traditional success rate with the average number of system-user turns.

[1] For a summary of inspiring visualizations and an overview of current trends, please see: http://selection.datavisualization.ch

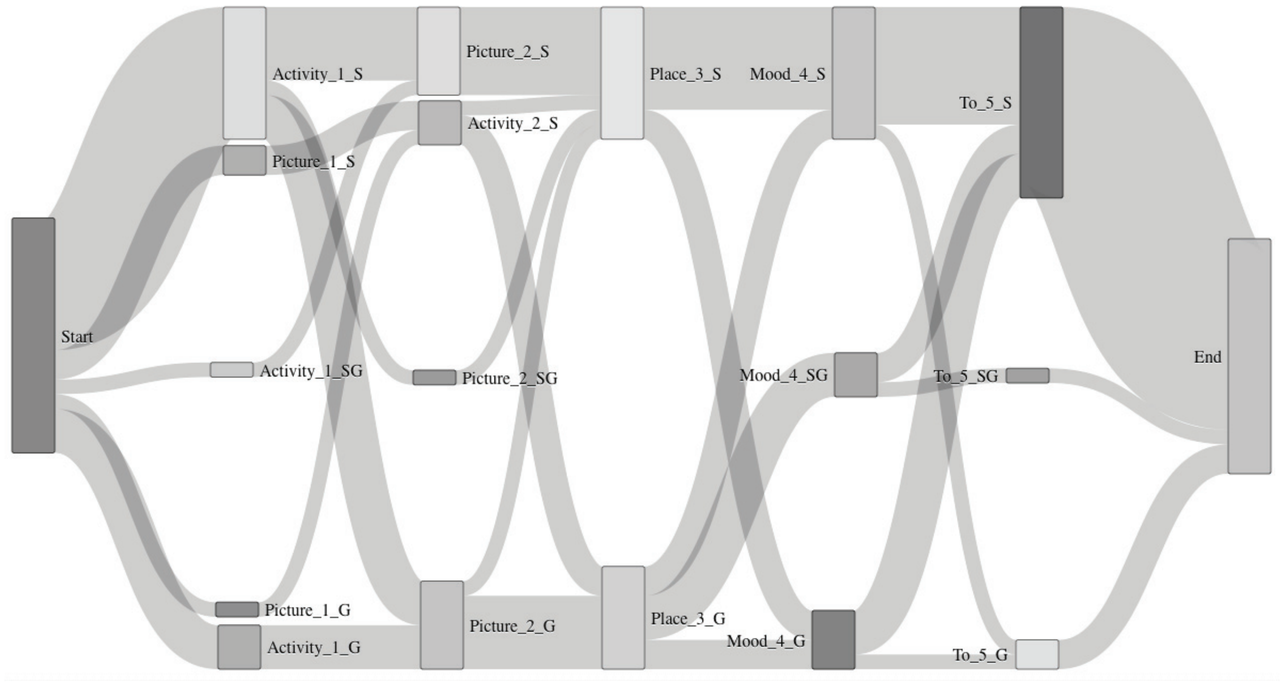

Figure 1: Visualization of multimodal dialogues for the experience sharing mockup application (labels consist of the semantic unit title, the ordinal number for the dialogue turn, and tags S for speech modality, G for gesture or SG for speech and gesture).

Visualizing system performance

As new visualization tools become more widely available and used by more and more disciplines, the hurdle to experiment with data is becoming lower and lower. Figure 1 presents a way to visualize multimodal interaction.

The experience sharing application The data visualized was collected through a user study, implemented as a Wizard-of-Oz (WoZ) mockup of an experience sharing application. The user's task was to compile a daily summary of their activities and share with others by specifying 5 items: Activity, Picture, Place, Mood, and To Whom to send the summary. The modality inputs were speech, gesture, and their combinations, based on which the wizard controlled the flow of the interaction. A total of 16 users participated in the study (10 females, 6 males). Two subjects were under age 30; 4 were between 30 and 40; and 10 of the subjects were 40 old or older. Out of the 16 participants, 7 had used Automatic Speech Recognition before, and only 3 subjects reported some prior experience with gesture inputs.

Implementation The visualization belongs to the class of Sankey diagrams, originally created for visualizing the magnitude of connections and the flow between the nodes of a network. Here, the dialogue paths represent the flow, while the dialogue states correspond to the nodes. The implementation utilizes D3.js[2], by Mike Bostock, and the Sankey plugin, developed by Jason Davies and Mike Bostock.

Looking into the details The basic structure follows a left-to-right flow with the start and end states on the respective ends. The five dialogue steps are placed horizontally between the end states. The notation for each state indicates the semantic unit acquired in that state, the ordinal number for the turn in the dialogue path (1-5), and a tag that refers to the modality used (S - speech, G - gesture, SG - speech and gesture). For easier interpretation of the visualization, the placement of states is organized so that states with speech modality are in the upper part of the visualization, states with gesture are in the bottom, and the truly multimodal states are in the middle. The paths connecting the states indicate turn-taking. A horizontal path refers to no change in the modality used, while a diagonal movement indicates the user switched from one modality to another.

The width of the paths between the individual states indicate the proportion of use of a certain modality relative to all possible paths for that state. For instance, ca. 2/3 of the users chose the speech modality in the first step (11 out of 16) to select Activity or Picture, 1/4 chose the gesture (4 out of 16), and only 1 chose the truly multimodal interaction.

The visualization provides interactivity to explore more details: by hovering over the paths, a tooltip is displayed (not shown in the figure) with the basic statistics for that particular turn and the verbal expressions most frequently used. For a dialogue state, the statistics include all incoming and outgoing transitions and the relative use of modalities.

By design, to achieve a trade-off between complexity and the goals of the study, the users did not have complete control to direct the dialogues. Only in the first two turns could

[2] Data-Driven Documents Javascript library, http://d3js.org

they choose freely between Activity or Picture, after which the order of semantic items was fixed. Also, no recognition or modality integration errors were simulated, thus all paths ended in the End state.

From the developer's and designer's perspective, the most important aspects of the visualization are to see the big picture, and to discover if problems occur. For instance, identifying if a multimodal transition always breaks in a given state - that is, how the Multimodal Integration performs - is of paramount interest. The designer can also see if users prefer only certain modalities over others, as in this case speech over gesture and only minimal multimodal inputs, and how the current implementation matches the desired functionalities. The tool provides both high level and more detailed views to achieve this.

4 Conclusions and Future Work

The paper presented our work on the evaluation process of Multimodal Dialogue Systems. The focus was component level performance evaluation, as well as on interactive visualization of multimodal interactions.

Regarding future work, we plan to continue the development of the statistical Multimodal Integration method with a better equipped evaluation framework. Our ultimate goals are to enable the use of *EMMA-like log files as input* to the visualization tool and with an *interactive web-based solution* to help designers and developers gain insights more easily from the dialogue state level, e.g. uni- and bigrams of the used verbal expressions, as well as to the overall application level.

5 Acknowledgement

We are thankful to all the participants in the user study in Helsinki and in Palo Alto. Also, we want to thank the reviewers for their valuable suggestions.

References

Adler, A., and Davis, R. 2007. Speech and sketching for multimodal design. In *ACM SIGGRAPH 2007 courses*, SIGGRAPH '07. New York, NY, USA: ACM.

Boda, P. P. 2000. Visualisation of spoken dialogues. In *Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH 2000, Beijing, China, October 16-20, 2000*, 146–149. ISCA.

Boda, P. P. 2004. Multimodal interaction in a wider sense. In *Invited talk at COLING 2004 Satellite Workshop on Robust and Adaptive Information Processing for Mobile Speech Interfaces*, COLING 2004.

Boda, P. P. 2006. A contextual multimodal integrator. In *Proceedings of the 8th international conference on Multimodal interfaces*, ICMI '06, 129–130. New York, NY, USA: ACM.

Dumas, B.; Ingold, R.; and Lalanne, D. 2009. Benchmarking fusion engines of multimodal interactive systems. In *Proceedings of the 2009 International Conference on Multimodal Interfaces and Machine Learning for Multimodal Interaction*, ICMI-MLMI '09, Cambridge, Massachusetts, USA, 169–176. New York, NY, USA: ACM.

Flippo, F.; Krebs, A.; and Marsic, I. 2003. A framework for rapid development of multimodal interfaces. In *Proceedings of the 5th international conference on Multimodal interfaces*, ICMI '03, 109–116. New York, NY, USA: ACM.

Fraser, N. M. 1993. The SUNDIAL speech understanding and dialogue project: results and implications for translation. In *ASLIB Proceedings*, volume 46, issue 5, 141 – 148. London, UK, UK: Aslib, The Association for Information Management.

Hakulinen, J.; Turunen, M.; and Salonen, E.-P. 2005. Visualization of spoken dialogue systems for demonstration, debugging and tutoring. In *In Proceedings of Interspeech'2005 - Eurospeech 9th European Conference on Speech Communication and Technology*, 853–856.

Johnston, M., and Bangalore, S. 2000. Finite-state multimodal parsing and understanding. In *Proceedings of the 18th conference on Computational linguistics - Volume 1*, COLING '00, 369–375. Stroudsburg, PA, USA: Association for Computational Linguistics.

Johnston, M.; Baggia, P.; Burnett, D. C.; Carter, J.; Dahl, D. A.; McCobb, G.; and Raggett, D. 2009. EMMA: Extensible MultiModal Annotation markup language. http://www.w3.org/TR/emma/.

Johnston, M. 1998. Unification-based multimodal parsing. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, COLING-ACL '98, 624–630. Stroudsburg, PA, USA: Association for Computational Linguistics.

Lalanne, D.; Nigay, L.; Palanque, P.; Robinson, P.; Vanderdonckt, J.; and Ladry, J.-F. 2009. Fusion engines for multimodal input: a survey. In *Proceedings of the 2009 International Conference on Multimodal Interfaces and Machine Learning for Multimodal Interaction*, ICMI-MLMI '09, Cambridge, Massachusetts, USA, 153–160. New York, NY, USA: ACM.

Simpson, A., and Fraser, N. M. 1993. Black box and glass box evaluation of the SUNDIAL system. In *EUROSPEECH 1993*, 1423–1427. ISCA.

Tat, A., and Carpendale, M. S. T. 2002. Visualising human dialog. In *IV*, 16–21.

Walker, M. A.; Litman, D. J.; Kamm, C. A.; and Abella, A. 1998. Evaluating spoken dialogue agents with PARADISE: Two case studies. *Computer Speech & Language* 12(4):317–347.

Wollmer, M.; Al-Hames, M.; Eyben, F.; Schuller, B.; and Rigoll, G. 2009. A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams. *Neurocomputing* 73(1-3):366–380.

Wu, L.; Oviatt, S. L.; and Cohen, P. R. 1999. Multimodal integration - a statistical view. *IEEE Transactions on Multimedia* 1:334–341.

Yau, N. 2011. *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*. John Wiley & Sons.