Ein dreistufiger Ansatz zur Evaluation von audio-visueller Systemausgabe

Christine Kühnel, Benjamin Weiss, Sebastian Möller

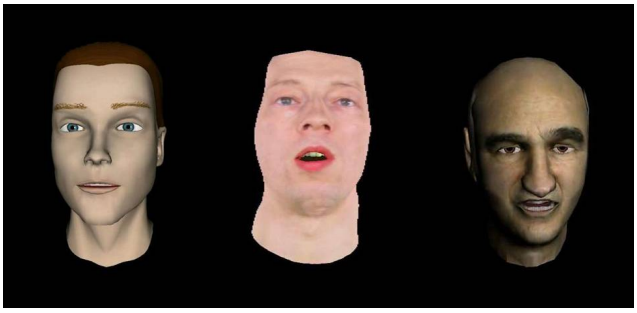
Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin, Email: christine.kuehnel@telekom.de

Einleitung

Die Vorzüge der audio-visuellen Sprachausgabe - wie zum Beispiel ein verbessertes Hörverstehen - haben zu einer verstärkten Anwendung von sogenannten 'Sprechenden Köpfen' geführt. Diese werden zum Beispiel zum Erlernen von Fremdsprachen oder in der Logopädie eingesetzt. Weiterhin wird untersucht, ob Sprechende Köpfe Schwerhörige beim Telefonieren unterstützen können und es besteht die Hoffnung, dass Sprechende Köpfe als Systemausgabe die Mensch-Maschine Kommunikation verbessern werden. Trotz dieses breiten Einsatzfeldes gibt es neben der Überprüfung des Textverstehens wenige oder keine standardisierten Evaluationsmethoden. Hier setzt unsere Arbeit an. In diesem Beitrag wird ein dreistufiger Ansatz zur Evaluierung von Sprechenden Köpfen vorgestellt. Bei dieser Methodik werden die zu evaluierenden Sprechenden Köpfe zunächst einem passiven Nutzer als Videos präsentiert und von diesem bewertet. In einem zweiten Schritt erfolgt die Bewertung der Sprechenden Köpfe im Anschluss an eine - über ein Wizard-of-Oz-Experiment simulierte - Interaktion in einer Laborumgebung. Und schließlich bewerten die Nutzer den Sprechenden Kopf als Ausgabekomponente des Endsystems, nachdem sie aufgabenbasiert mit dem System interagiert haben. Jeder einzelne Schritt dieser Methodik wird beispielhaft anhand einer bereits durchgeführten Evaluation im Smart-Home-Bereich erläutert. Zudem wird ein kritischer Überblick über die erzielten Ergebnisse gegeben.

Schritt Eins - Qualitätsaspekte Sprechender Köpfe

Um eine Referenzmessung von sechs Sprechenden Köpfen zu erhalten wurde zunächst ein passives Testszenario gewählt. Die eingesetzten sprechen Köpfe sind Kombinationen aus zwei verschiedenen TTS-Modulen und drei visuellen Synthesemodulen (siehe Abb. 1), darunter der HeadZero aus dem Thinking-Systems-Projekt (als deutsche Version [1]). Die Teilnehmer bewerteten audiovisuelle Aufzeichnungen der Sprechenden Köpfe, die verschiedene einzelne Sätze aus dem Smart-Home Bereich artikulierten. In einer zweiten Version wurde eine Auswahl der Sätze zu Blöcken zusammengefasst, um einen stärkeren Eindruck der Sprechenden Köpfe zu erreichen. Verständlichkeit wurde nicht überprüft; die Bewertungsskalen umfassten Sprachqualität, visuelle Qualität und Gesamtqualität, sowie ein Semantisches Differential zum Erfassen hedonischer Aspekte wie Sympathie und Attraktivität. Die im Folgenden präsentierten Ergebnisse können nicht verallgemeinert werden, da nicht alle Aspekte mit einem vollfaktoriellen Design überprüft werden konnten. Trotzdem konnte gezeigt werden, dass diese Methode gut geeignet ist Qualitätsunterschiede für Kopf- und Sprachmodule zu erfassen: Die Teilnehmer unterschieden deutlich zwischen Sprachqualität und visueller Qualität, die sich wiederum auf die Gesamtqualität auswirken [3]. Durch die Analyse des Semantischen Differentials konnten perzeptive Aspekte identifiziert werden, die ebenfalls einen Einfluss auf die Gesamtqualität haben: Wie realistisch die Sprechenden Köpfe sind spielt eben so wenig eine Rolle, wie gegebenenfalls randomisiert auftretende Veränderungen des Gesichtsausdrucks. Dagegen weist die wahrgenommene Attraktivität die stärkste Korrelation mit der Gesamtqualität auf [6]. Vergleichbare Ergebnisse konnten in einem Webexperiment erzielt werden [7].



(a) MASSY (b) Speaker Cloning (c) Thinking Head

Abbildung 1: Drei Kopfmodule.

Schritt Zwei - Qualitätserfassung in Interaktion

Ein deutlicher Unterschied besteht zwischen den Evaluationstudien, in denen die Teilnehmer mit den Sprechenden Köpfen (oder Embodied Conversational Agents, ECA) interagieren und den Studien, in denen keine Interaktion stattfindet. Insbesondere in Tutorsystemen, eingesetzt zur Unterstützung der Lehre, findet häufig nur eine einseitige Kommunikation statt: der Tutor übermittelt Wissen an den Nutzer. Diese Unterscheidung hat sich als wichtig herausgestellt, da zum Beispiel der menschlichere ECA im nicht-interaktiven Szenario als intelligenter empfunden wurde, aber nicht nach einer Interaktion [2]. Um den Einfluss von Interaktivität zu überprüfen wurde ein weiterer Test als WOZ-Experiment durchgeführt: Mit Hilfe des ECAs mussten die Teilnehmer kurze Aufgaben in einem simulierten Smart-Home-Szenario lösen, darunter die Abfrage des elektronischen Fernsehprogramms. Die Ergebnisse unterscheiden sich deutlich von denen erzielt im passiven Szenario: Die ursprünglich festgestellten Unterschiede zwischen den Köpfen finden sich nur mehr auf einigen der erfassten Skalen, nicht aber in der Be-

Abbildung 2: Das Wohnzimmer, in dem das Smart-Home-System installiert ist.

wertung der Gesamtqualität. So haben wahrgenommene Unterschiede in der Qualität der Laut- und Lippensynchronisation in diesem Szenario keinen Einfluss auf die Gesamtqualität [5]. Zusammenfassend schließen wir, dass der Grad an Interaktivität einen hochrelevanten Einfluss auf die Wahrnehmung und / oder den inneren Beurteilungsprozess hat. Ob diese Unterschiede zwischen den Ergebnissen des passiven und des interaktiven Szenarios aus der Ablenkung oder der Vermischung von Systemqualität und Qualität des Sprechenden Kopfes resultieren bleibt offen. Fest steht: Bewertungen, die in einem passiven Szenario erzielt wurden, können nicht auf ein interaktives System übertragen werden.

Schritt Drei - Umgebungseinflüsse

Über die präsentierten Untersuchungen hinausgehend sind wir interessiert an Einflüssen der Umgebung auf die Bewertungen. So führt zum Beispiel die Darstellung von geeigneter redundanter Information (in diesem Fall Programminformationen und Listen von aufgezeichneten Filmen) auf einem zusätzlichen Bildschirm zu einer Steigerung der Gesamtqualität des Sprechenden Kopfes [5]. Die deutet darauf hin, dass die Sprechenden Köpfe als Repräsentation des Gesamtsystems aufgefasst und entsprechend bewertet werden. Abschließend wurde noch der 'Persona Effect'– also der wahrgenommene oder tatsächliche positive Einfluss eines ECAs auf die Interaktion – in einem realen Wohnzimmer (siehe Abb. 2) untersucht. Die Teilnehmer saßen nicht in einem Labor sondern auf einem gemütlichen Sofa. Wie auf Basis der bereits dargestellten Ergebnisse angenommen, wurden das System mit reiner Sprachausgabe und das System mit Sprechendem Kopf gleich bewertet [2]. Gründe hierfür könnten in der Funktionalität des Smart-Home-Systems zu finden sein: Die Steuerung der Geräte erfordert keinen komplexen Dialog. Tatsächlich kommt die sprachliche Systemausgabe fast nur in Fällen von Kommunikationsfehlern zum Einsatz. Es ist nicht ausgeschlossen, dass in anderen Bereichen mit intensiverem Dialog die verwendeten Sprechenden Köpfe einen 'Persona Effect' auslösen können.

Schlussfolgerungen

Bei der Erfassung der Qualität von Sprechenden Köpfen und ihres Einsatzes für eine spezifische Applikation erlaubt der vorgestellte Ansatz die Identifikation wichtiger Kontextfaktoren: der Grad an Interaktion sowie der Einfluss der Umgebung. Wir empfehlen entsprechend bei der Wahl oder Entwicklung eines ECAs zunächst die Qualität der Module in einem passiven Szenario bewerten zu lassen. Dies ist auch als Online-Test möglich und erlaubt eine schnelle und kostengünstige Auswahl und Kombination von äußerlichem Erscheinungsbild und Stimme, sowie zusätzlicher Eigenschaften wie der Generierung von Gesichtsausdrücken. Sobald der geeignete ECA ausgewählt wurde ist aber eine Evaluation in Interaktion unabdingbar, am besten in der Umgebung, in der der ECA letztlich zum Einsatz kommt, und mit realistischen Aufgaben. Die Vielzahl der möglichen Einflussfaktoren ist zu groß, als dass auf den letzten Schritt verzichtet werden könnte.

Literatur

[1] S. Fagel, C. Kühnel, B. Weiss, I. Wechsung, and S. Möller: A comparison of German talking heads in a smart home environment. In Workshop on Audio-Visual Speech Processing (AVSP), Togalooma, 2008.

[2] C. Kühnel, B. Weiss, and S. Möller: Talking heads for interacting with spoken dialog smart-home systems. In INTERSPEECH, Brighton, pages 304-307, 2009.

[3] C. Kühnel, B. Weiss, I. Wechsung, S. Fagel, and S. Möller: Evaluating talking heads for smart home systems. In 10thInternational Conference on Multimodal Interfaces (ICMI), Chania, pages 81-84, 2008.

[4] D. Massaro, M. Cohen, J. Beskow, and R. Cole: Developing and evaluating conversational agents. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, Embodied conversational agents, pages 286-318. MIT Press, 2000.

[5] B. Weiss, C. Kühnel, I. Wechsung, S. Fagel, and S. Möller: Quality of talking heads in different interaction and media contexts. Accepted for Speech Communication.

[6] B. Weiss, C. Kühnel, I. Wechsung, S. Möller, and S. Fagel: Comparison of different talking heads in non-interactive settings. In Proc. Human Computer Interaction International (HCII), San Diego, pages 349-357, July 2009.

[7] B. Weiss, C. Kühnel, I. Wechsung, S. Möller, and S. Fagel: Web-based evaluation of talking heads: - how valid is it? In 9th International Conference on Intelligent Virtual Agents (IVA), Amsterdam, page 552. 2009.