LIPPS – Eine dialogbasierte, audio-visuelle Trainingshilfe für das Absehen zum Einsatz mit Schwerhörigen

Hermann Gebert[1], Hans-Heinrich Bothe[2]

[1] *Technische Universität Berlin, 10623 Berlin, Deutschland, Email: h.gebert@xlp.de*

[2] *Technische Universität Dänemark, 2800 Kgs. Lyngby, Dänemark, Email: hhb@ieee.org*

Einführung

Die Kommunikationsmöglichkeiten von hörgeschädigten Personen im Umgang mit nicht speziell ausgebildeten Normalhörenden sind stark begrenzt. Eine Möglichkeit zur Verständigung liegt im Absehen der gesprochenen Wörter von den Lippen und vom Gesicht des Kommunikationspartners. Das Absehen ermöglicht jedoch nur eine geringe Erkennungsrate von ungefähr 30%. Die Erkennungsrate kann gesteigert werden, wenn die absehende Person den Kontext bzw. das Thema kennt und über ein ausgezeichnetes Sprachwissen verfügt. Das Erlernen des Absehens bedarf einer fundierten Ausbildung und stetigen Trainings. Das vorgestellte System kann zum Erlernen und Üben des Absehens im Klassenverbund oder im privaten Umfeld genutzt werden. Es soll keinen Ersatz, sondern vielmehr eine Ergänzung zum Absehunterricht darstellen.

Aufbau

Das vorliegende System orientiert sich an dem Aufbau eines Sprachdialogsystems und verfolgt einen modularisierten Aufbau.

Struktur

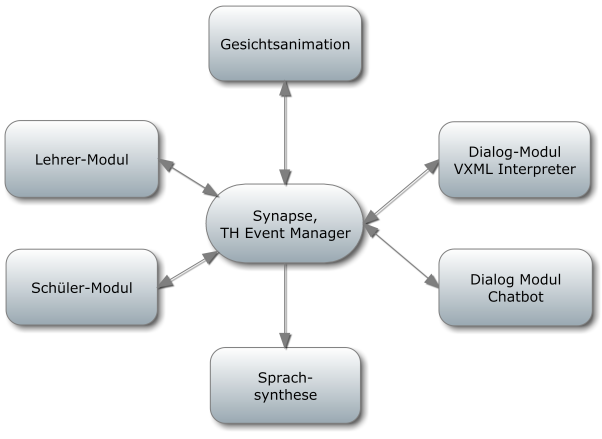

Abbildung 1: Struktur und Kommunikationspfade der genutzten Komponenten der vorliegenden Lernhilfe.

Die zentrale Datenverwaltung ist verbunden mit den neu entwickelten Lehrer- und Schüler-Modulen, welche die Benutzerschnittstelle darstellen und für die Dateneingabe genutzt werden. Die Dialog-Module setzen die ausgewählten und später dargestellten pädagogischen Modelle um und sind für den Dialogverlauf zuständig. Die audio-visuelle Ausgabe erfolgt durch die Module für die Sprachsynthese und die Gesichtsanimation.

Pädagogisches Modell

Wiederholung

Die für die jeweilige Lektion ausgewählten Lerneinheiten werden mehrmals präsentiert, damit der Schüler die Möglichkeit hat sich die Bewegungen einzuprägen. Die Antworten der jeweiligen Lektion werden so oft dargestellt, wie vom Lehrer definiert wurde. An dieser Stelle erfolgt noch keine Abfrage des Gelernten.

Geführter Dialog

Der geführte Dialog stellt die Möglichkeit dar, einzelne Wörter, kurze Sätze mit Auslassungen, sowie Minimalpaare und Vokal- bzw. Silbenzählungen zu modellieren. Die Dialoggestaltung ist dem ausgebildeten Pädagogen überlassen und wird durch einen VoiceXML Interpreter abgearbeitet. Ein frei veränderbares Templatesystem garantiert große Flexibilität. Der Lehrer hat die Möglichkeit zwei verschiedene Grammatiken anzugeben, welche die schlechte Unterscheidbarkeit von bestimmten Phonemen kompensieren sollen. So ist die visuelle Wahrnehmung des Wortes “Mama” identisch mit der des Wortes “Papa”. Bei unklarem Kontext müssten somit beide Antworten als korrekt gewertet werden. Die zwei Grammatiken erlauben eine zusätzliche Unterscheidung. Um den Schüler beim Erlernen der einzelnen Einheiten zu unterstützen, wurde eine Lerntechnik integriert, welche eine Wiederholung in definierten Zeitabständen ermöglicht (Spaced Repetition). Das System ist allgemein als *Lernkartei* oder *Leitner System* bekannt [2].

Freier Dialog

Wenn der Schüler bereits das Absehen beherrscht und einen fundierten Wortschatz besitzt, kann er den freien Dialog nutzen. Der freie Dialog basiert auf dem A.L.I.C.E. Chatbot [3], welcher bestimmte Regeln zur Mustererkennung nutzt um einen möglichst realistischen und natürlichsprachlichen Dialog zu erzeugen. Des Weiteren verfügt das System über ein Gedächtnis, was zu einem noch realistischeren Dialog führt und dem Schüler das Üben von Alltagsgesprächen ermöglicht.

Szenen-Erstellung

Neben der Verwaltung der Schüler und der Erstellung von Lektionen & strukturierten Lerneinheiten, ist die Möglichkeit individuelle Szenen zu gestalten, eine weitere

Besonderheit des vorliegenden Systems. In Verbindung mit der genutzten Gesichtsanimation [1] ist es möglich eine Rotation, sowie eine Translation des Gesichtes auf allen drei Achsen durchzuführen. Des Weiteren existieren zwei unabhängige Lichtquellen, welche in ihrer Streuung, Lichtstärke und Position variiert werden können. Das System erlaubt zusätzlich eine Größenanpassung des Gesichtes und die Integration von individuellen Hintergrundbildern. Die auditive Szene kann durch das Einfügen von Hintergrundgeräuschen und Rauschen beeinflusst werden. Ferner ist die Möglichkeit gegeben, eigene Gesichter und Accessoires, sowie Stimmen zu nutzen.

Evaluierung

Die Evaluierung des Systems bezieht sich ausschließlich auf die Komponenten: Wiederholung, geführter Dialog und Gesichtsanimation, wobei das Leitner System im geführten Dialog keine Beachtung findet.

Testaufbau

Der Test ist unterteilt in einen audio-visuellen und einen rein auditiven Test, welche jeweils eine Stunde dauern. Er wurde über einen Zeitraum von fünf Tagen mit äquidistanten Pausen durchgeführt. Der Test präsentiert drei mal fünfzehn Zahlen, welche zufällig einem gemeinsamen Wortkorpus entnommen wurden, welcher Zahlen aus dem Bereich 21 - 99 erhält. Die Zahlen werden nur an dem jeweiligen Testtag genutzt und randomisiert vorgetragen. Um eine Hörbeeinträchtigung zu simulieren, wurde das Sprachsignal mit einem *speech-shaped noise* in dem Maße überlagert, dass die Versuchspersonen den Inhalt nicht mehr verstehen konnten. Das so ermittelte Signal-Rausch-Verhältnis wird jeweils um zwei und vier dB verstärkt bzw. abgeschwächt und in zufälliger Reihenfolge eingespielt. Der Proband hat die Aufgabe, die gesagte Zahl von den Lippen abzusehen und in die Testumgebung einzugeben. Der rein auditive Test bietet keine Gesichtsanimation, liefert keine Rückmeldung an den Probanden und wurde nur am ersten und letzten Tag durchgeführt.

Auswertung

Beim rein auditiven Test werden die Eingaben direkt mit den nachgefragten Zahlen verglichen und bei Übereinstimmung als 100% korrekt angesehen. Zur Überprüfung der Adaption an das Rauschsignal ist nur die Differenz zwischen dem ersten und letzten Tag von Bedeutung. Beim audio-visuellen Test wurden die genannten Zahlen und die Eingaben des Probanden zuerst phonetisch transkribiert bevor eine Übersetzung in die repräsentativen Viseme erfolgte. Die so gewonnenen Visemfolgen wurden zur einfacheren Verarbeitung kodiert und mit Hilfe eines Algorithmus zur Berechnung der Levenshtein-Distanz verglichen. Hierbei wurden Auslassungen, Vertauschungen und Einfügungen berücksichtigt und die Ähnlichkeit der zwei Folgen bestimmt. Zur einfacheren Interpretation wurden die Erkennungsraten der Versuchspersonen arithmetisch gemittelt und durch eine lineare Regression der Trend bestimmt. Abbildung 2

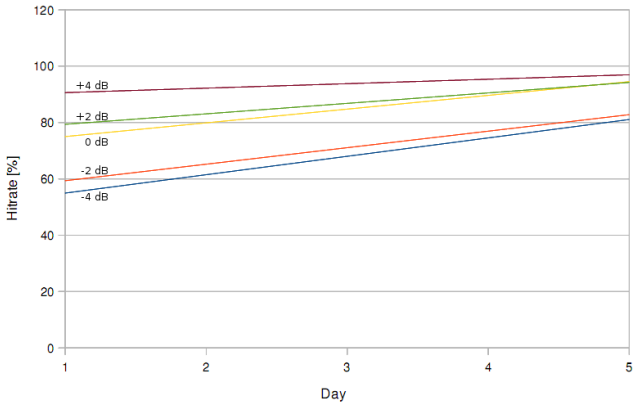

Abbildung 2: Arithmetischer Mittelwert der audio-visuellen Erkennungsrate aller Probanden über der Zeit

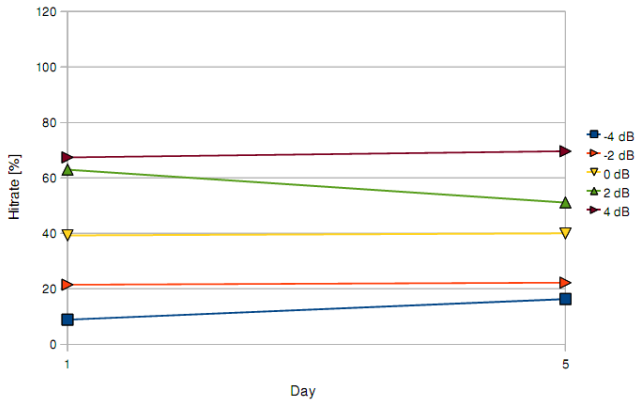

Abbildung 3: Arithmetischer Mittelwert der rein auditiven Erkennungsrate aller Probanden über der Zeit

lässt erkennen, dass die Absehfähigkeiten der Probanden innerhalb der Testperiode stark gestiegen sind. Im Schnitt handelt es sich um eine Erhöhung von 21%. Die individuellen Werte reichen von 13,73% bis hin zu 34,22%. Abbildung 3 unterstützt die Vermutung, dass die Erhöhung der Erkennungsrate nicht oder nur sehr gering mit einer Anpassung an das Rauschsignal zusammenhängt. Weiterhin ist zu erkennen, dass die Steigerung bei geringem S/N-Verhältnis steiler ist als bei einem hohen. Dennoch wird ebenfalls die Verständlichkeit bei geringem Rauschen erhöht. Dies lässt darauf schließen, dass die Nutzung der Trainingshilfe, bei den verschiedensten Stärken einer Hörschädigung, zu einer Steigerung der Absehfähigkeit führen kann.

Literatur

[1] Luerssen, M; Lewis, T: Head X: Tailorable audiovisual synthesis for ECAs. Interacting with Intelligent Virtual Characters Workshop (IIVC), Sydney, Australia, Dec 2009

[2] Leitner, S.: So lernt man lernen. Herder Verlag, Freiburg, 1973

[3] Wallace, R. S.: A.L.I.C.E. Artificial Intelligence Foundation, URL: `http://www.alicebot.org`