Ein Framework für die indirekte Bewertung der Qualitätswahrnehmung audiovisueller Sprache

Stephen Wilson

Quality and Usability Laboratory, Deutsche Telekom Innovation Laboratories
Ernst-Reuter-Platz 7, 10587 Berlin, E-Mail:stephen.wilson@telekom.de

Einleitung

Die Multimodalität in der Informations- und Kommunikationstechnik hat in den letzten Jahren eine große Signifikanz erworben. Viele Systeme sind jetzt auf veschiedene Input- und Output-Modalitäten basiert, (z.B. Text, Sprache, Antasten) und viele Online-Services liefern multimodale Inhalt ab, (z.B. IP-TV, Videotelefonie). Die Qualitätswahrnehmung multimodaler Systemen und Services ist, folglich, sehr wichtig geworden, und es gibt viele Faktoren die darauf einwirken können, (z.B. Störsignale, Bildstandsschwankungen usw.).

Für multimodale Systeme, die auf audiovisuelle Sprache basiert sind, ist die Synchronität zwischen Ton und Bild eine wichtige Einflußgröße hinsichtlich der Qualitätswahrnehmung und infolgedessen der Akzeptanz des Systems. Dieser Beitrag stellt ein Framework vor, welches darauf abzielt mittels indirekter Methoden eine automatische Bewertung der wahrgenommen Qualität sowie der Akzeptanz asynchroner audiovisueller Sprache vorzunehmen.

Ziele des Frameworks

Das Framework hat folgende Ziele:

- den Effekt verschiedener audiovisueller Asynchronitätsstufen auf Qualitätswahrnehmung und Akzeptanz zu erfassen,
- zuverlässige Verfahren zur automatischen Messung der Asynchronität zwischen Ton und Bild zu untersuchen,
- ein robustes Modell zur indireketen Bewertung der Qualitätswahrnehmung sowie der Akzeptanz audiovisueller Sprache zu entwickeln.

Gliederung des Frameworks

1. Direkte Bewertungen von der Qualitätswahrnehmung audiovisueller Sprache versammeln
2. Die Asynchronität zwischen Ton und Bild automatisch messen
3. Verbindungen zwischen die direkt gemessenen Bewertungen von 1. und die automatisch gemessenen Asynchronitätswerten von 2. automatisch lernen, was ein „erstes“ Qualitätsmodell ergibt.
4. Ein halbautomatisches iteratives Verfahren verwenden, um das Qualitätsmodell zu testen und robuster zu machen.

Audiovisuelle Daten

Daten von der CUAVE Datenbank wurde als audiovisuelle Stimuli benutzt [1].

Direkte Bewertung der Qualitätswahrnehmung

Jedes indirekte Bewertungsmodell muß im Prinzip auf direkt gesammelte Bewertungen basiert sein. Solche direkte Bewertungen können nur durch Experimente gemessen werden. In dem vorgestellten Framework werden solche direkte Bewertungen verallgemeinert und als eine Basis benutzt, worauf ein indirektes Qualitätsmodell entwickelt wird. Verschiedene audiovisuelle Stimuli mit unterschiedlichen Asynchronitätsstufen wurden vorbereitet [1]. Die benutzte Stufen waren auf ITU-R BT.1359 basiert [2], ein Teil von welchen in Tabelle 1 zusammengefaßt ist.

Tabelle 1: Zusammenfassung der Asynchrontitätsstufen

Audioverzögerung (ms)	Videoverzögerung (ms)
95	22,5
125	45
155	67,5
180	90
310	135
435	180

Extrahierung von Audiovisuellen Merkmalen

Visuelle Merkmale

Ein allgemeines Lippe-Modell wurde mit dem „open source“ Active Appearance Model API, (AAM-API) [3], entwickelt und als einen Lip-Tracker verwendet. Active Appearance Models sind verallgemeinerbare statistische Modelle die sich von Form sowie Graustufenwerten charakterisieren, [4]. Der Lip-Tracker wird benutzt um einen normalisierten Lippenbereich zu extrahieren, den durch eine Diskrete Cosinus-Transformation (DCT) in einen 30-dimensionalen Vektor umgewandelt ist. Ableitungen der ersten und zweiten Ordnungen werden auch berechnet und dazu verkettet, was denn einen 90-dimensionalen visuellen Merkmalsvektor ergibt.

Akustische Merkmale

Mel-frequency Cepstral Coefficients (MFCC) sind weitverbreitet in der Sprachwissenschaft und

Sprachtechnologie. Die ersten 12 MFCC sowie die Signalenergie werden jede 10ms berechnet, was einen 13-dimensionalen Vektor ergibt. Die Ableitungen der ersten und zweiten Ordnungen werden berechnet und dazu verkettet und ergibt einen 39-dimensionalen akustischen Merkmalsvektor.

Abtastrate

Die visuelle Merkmale haben eine Abtastrate von 29.87fps, die akustische eine von 100Hz. Eine Geradeninterpolation wird benutzt, um die visuelle Abtasterate mit der akustischen gleichzusetzen, damit beide eine Rate von 100Hz haben.

Abbildung 1: Output des AAM-Liptracking

Automatische Messung von audiovisueller Asynchronität

Ein statistisches Verfahren namens *Co-inertia Analysis* wird benutzt, um die Asynchronität automatisch zu messen. Co-inertia analysis stammt ursprünglich von der Ökologie und berechnet eine Maximalisierung von der Kovarianz zwischen zwei Datensätzen. Laut Rúa et al [6] wird so eine Analyse auch in der Sprachtechnologie verwendet, als eine Messung von audiovisueller Asynchronität.

Nachdem die audiovisuellen Merkmalen vom Signal extrahiert worden sind, werden sie als Input benutzt, um die Asynchronität zwischen Ton und Bild zu messen, durch eine Anwendung von *Co-inertia Analysis*, was ein „Synchronitätswert“ ergibt.

Automatisches Lernen von Verbindungen zwischen den direkt Bewertungen und den automatisch gemessenen Asynchronitätswerten

Die direkt gemessen Bewertungen sowie die automatisch gemessenen Asynchronitätswerten werden als Input zu einem Lernensablauf gegeben, wodurch die Verbindungen zwischen den beiden automatisch gelernt werden können. Als Output kommt ein „erstes“ Qualitätsmodell. Das Modell wird als ein „erstes“ bezeichnet, da es muß durch ein halbautomatisches iteratives Verfahren verbessert wird.

Halbautomatisches iteratives Verfahren zu der Verbesserung des Modells

Um das Modell robuster zu machen, wird es durch folgendes iteratives Ablauf geführt:

1. Neue audiovisuelle Stimuli werden automatisch von dem Qualitätsmodell, sowie direkt von neuen Versuchspersonen bewertet.
2. Die zwei Bewertungen werden miteinander verglichen. Wenn die beide miteinander nicht zustimmen, muß die Qualitätsmodells Bewertung „per Hand“ adjustiert werden, (wodurch das Verfahren halbautomatisch bezeichnet ist).
3. Die neuen adjustierten Daten werden als Input zu einem neuen Lernensablauf gegeben, damit das Modell robuster wird.

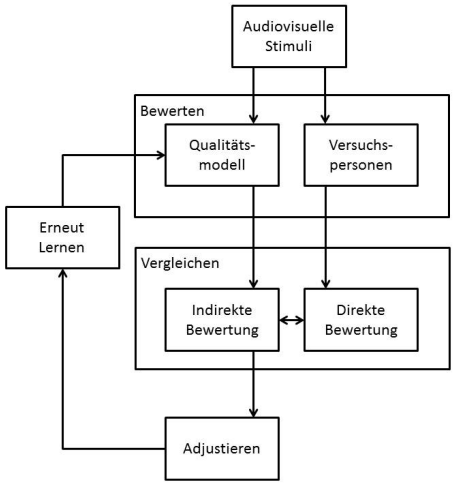

Abbildung 2: Überblick des halbautomatischen iterativen Verfahrens

Literatur

[1] E. K. Patterson and S. Gurbuz, Z. Tufekci and J. N. Gowdy. Moving-talker speaker-independent feature study and baseline results using the CUAVE multimodal speech corpus. EURASIP J. on App. Sig. Proc., 2002, 11, 1189–1201.

[2] ITU, http://www.itu.int/rec/R-REC-BT.1359/en, ITU-R BT.1359: Relative Timing of Sound and Vision for Broadcasting, 1998

[3] M. B. Stegmann, B. K. Ersbøll, R. Larsen: FAME -- A Flexible Appearance Modelling Environment, IEEE Transactions on Medical Imaging, IEEE, 2003

[4] T.F. Cootes, G.J. Edwards, C.J. Taylor: Active Appearance Models. J. IEEE Trans. Pattern Anal. Mac. Intell. Vol 23, No 6, pp 681-685, 2001

[5] S. Dolodec, D. Chessel. Co-Inertia Analysis: An alternative method for studying species-environment relationships. Freshwater Biology, 1994.

[6] E. A. Rúa, H. Bredin, C. García-Mateo, G. Chollet, and D. González-Jiménez, "Audio-Visual Speech Asynchrony Detection using Co-Inertia Analysis and CHMMs. J .Pattern Anal. and App. 2007