ÜBERLEGUNGEN ZU WAHRNEHMBAREN LÄCHELN IN SYNTHETISCHEN STIMMEN

Jürgen Trouvain[1], *Benjamin Weiss*[2]

[1]*Universität des Saarlandes,* [2]*Technische Universität Berlin*
trouvain@lst.uni-saarland.de, benjamin.weiss@tu-berlin.de

Kurzfassung: In dem vorliegenden Aufsatz beschäftigen wir uns mit der Frage, ob und falls ja, wie hörbares Lächeln für Sprachsynthese-Applikationen eingesetzt werden kann. In der Mensch-Mensch-Kommunikation kann Lächeln ganz verschiedenen Funktionen dienen, um beispielsweise Höflichkeit zu signalisieren, als Anzeichen von Vertrauenswürdigkeit und andere Aspekte, um Sprecher sympathisch wirken zu lassen. In der Mensch-Maschine-Kommunikation ist hörbares Lächeln weitestgehend unerforscht, könnte sich aber von Vorteil in verschiedenen Anwendungen wie Dialogsystemen und individualisierten Sprechprothesen erweisen. Eine große Herausforderung besteht darin das begrenzte Wissen der Details hörbar gelächelter Sprache für verschiedene Synthesetechniken nutzbar zu machen, aber auch wie Lächeln in Dialogen zu modellieren ist und wie man den Umgang von Nutzern mit diesem Detail testet.

1 Einführung

Soll man bei einer synthetischen Stimme erkennen können, dass sie lächelt? Sollen sprechfähige Maschinen in Form von Computern, virtuellen Agenten oder sozialen Robotern lächeln können? Falls ja, wie sollte ein solches Vorhaben angegangen werden?

Angenommen, dass diese Fragen zwar zunächst sonderbar erscheinen, aber nicht unwesentlich in der Forschung zur Mensch-Maschine-Kommunikation sind, dann besteht ein komplexer Klärungsbedarf, bevor man sich der technischen Umsetzung und der Überprüfung dieser Umsetzung durch Benutzer widmet. Die vorliegenden Überlegungen zu wahrnehmbarem Lächeln in synthetischen Stimmen widmen sich genau dieser Aufgabe. Dabei sollen Erkenntnisse, aber auch Probleme auf ganz verschiedenen Ebenen identifiziert, definiert und Lösungsansätze formuliert werden. Es soll sich bei diesen durchaus kritischen Überlegungen nicht um "wishful thinking" handeln, auch nicht um eine Machbarkeitsstudie, sondern um die Beleuchtung einer möglichen Bereicherung technischer Systeme im Sinne einer verstärkten Natürlichkeit in der Interaktion mit dem Benutzer.

In alltäglicher Interaktion spielt Lächeln eine sehr wichtige Rolle mit einer Vielzahl von Funktionen: als Anzeichen von Freundlichkeit und Höflichkeit, zum Ausdruck von Amüsiertheit und Erheiterung, zur Markierung von Ironie und Nichternsthaftigkeit, oder auch als Mittel zur Verführung. Oftmals wird Lächeln eingesetzt, um Vertrautheit zwischen Sprechern zu erzeugen, was auch eine große Auswirkung auf die soziale Attraktivität und Sympathie von Sprechern haben kann. In Face-to-Face-Situationen werden visuelle Signale als die primären angesehen, akustische Signale als sekundär. Beschränkt man sich jedoch auf den auditiv-akustischen Kanal, z.B. beim Telefonieren, ist Lächeln auch ausschließlich durch stimmliche und artikulatorische und damit akustische Merkmale als solches beim Sprechen wahrnehmbar.

Da in menschlicher Kommunikation Lächeln beim Sprechen hörbar ist und für zahlreiche affektiv-soziale Zwecke eingesetzt werden kann, stellt sich die Frage, ob nicht auch ähnliche Effekte in der Mensch-Maschine-Kommunikation erzielt werden könnten, indem man versucht Lächeln für synthetische Stimmen zu modellieren. Solche Fragen sind bislang nahezu unerforscht, könnten aber einen Mehrwert in verschiedenen technologischen Anwendungen

wie Dialogsystemen oder individualisierten Sprechprothesen erzeugen. Dabei sind die hier angestellten Überlegungen sind stets auf den auditiven Kanal ausgerichtet.

2 Funktionen von Lächeln

2.1 Affektiv-soziale Komponenten

Zu den prototypischen Assoziationen zu Lächeln zählen positive affektive Zustände wie Glück und Freude (z.B. wenn man ein schönes Geschenk bekommt), eine gute Stimmung (z.B. die Sonne nach mehreren grauen Tagen zu sehen) oder Zufriedenheit (z.B. wenn eine herausfordernde Aufgabe gemeistert wurde). Lächeln kann auch zur Verführung eingesetzt werden oder zum Ausdruck von Amüsiertheit und auch um Ironie zu markieren.

Die erwähnten positiv besetzten Arten von Lächeln haben einen authentischen Charakter. Zu den nicht-authentischen oder nicht echt gefühlten Arten des Lächelns gehören z.B. solche, bei denen negative Emotionen mit einem Ausdruck von Freude maskiert werden, oder bei denen man sich eigentlich unsicher, nervös oder peinlich berührt fühlt [16], aber auch solche, bei denen Überlegenheit gegenüber anderen signalisiert wird.

Der Unterschied zwischen "echtem" und "nicht-echtem" Lächeln manifestiert sich im visuellen Kanal hauptsächlich durch die Anspannung des Augenringmuskels (Duchenne-Lächeln) bzw. dessen Nicht-Anspannung [9]. Dieser Unterschied ist auch die Basis für den Unterschied zwischen vertrauenswürdigem und betrügerischem Verhalten [9]. Für den visuellen Kanal scheint diese fundamentale Unterscheidung etabliert, für den akustischen Kanal keineswegs.

Auf der Rezipientenseite gehört Lächeln zu den sozialen Signalen, die den Eindruck von und Interesse, Freundlichkeit und Intimität erzeugen [18, 4, 13]. In Bezug auf das Konzept sichtbarer Attraktivität verstärkt Lächeln die positive Bewertung weibliche Gesichter [21]. Zudem lächeln Frauen typischerweise beim Flirt [23]. Für männliche Gesichter ist dieser Effekt jedoch nicht so eindeutig [27,22]. Die Übertragung solcher Ergebnisse auf synthetische Stimmen gestaltet sich jedoch schwierig, da bei Fotomaterial natürlich kein, und bei Felddaten oft kein berichteter Bezug auf ko-verbales Lächeln vorhanden ist – und nur dieses ist hörbar.

2.2 Kulturelle Interpretationen

Es ist eine große Versuchung anzunehmen, dass positiv besetztes Lächeln stets als "authentisches" Lächeln mit einem Duchenne-Marker realisiert wird. Man könnte also annehmen, dass ein Lächeln einer bislang unbekannten Person als ein freundliches und positives Zeichen wahrgenommen wird und auch mit einer höheren Attraktivität verbunden wird (z.B. Fotos von Gesichtern bei Bewerbungen oder auf persönlichen Homepages). Es liegen jedoch Befunde vor, die zeigen, dass in nicht-westlichen Kulturen das authentische Lächeln nicht mit einem Duchenne-Marker verbunden wird [33]. Darüber hinaus zeigen Krys et al. [19] in einer groß angelegten kultur-übergreifenden Studie zur Interpretation von Gesichtsausdrücken mit Versuchspersonen aus mehr als 40 Ländern, dass in manchen Ländern lächelnde Gesichter unbekannter Personen einen negativen Eindruck hinterlassen. Dieser Effekt kann für hörbares Lächeln nicht ausgeschlossen werden, auch wenn dazu Studien zu fehlen scheinen.

Ein anderes Beispiel kultureller Verschiedenheit bezüglich des Lächelns bietet die Studie von Mui et al. [24]. Sie baten chinesische und niederländische Kindergartenkinder ein bestimmtes Spiel zu spielen, und das entweder alleine oder gemeinsam mit einem anderen Kind. Im Gegensatz zu den niederländischen Kindern, die zwischen beiden Bedingungen ihr Lächelverhalten nicht änderten, haben die chinesischen Kinder mehr gelächelt, wenn sie mit einem anderen Kind zusammen die Aufgabe bewältigten.

3 Akustische Merkmale gelächelter Sprache

Verschiedene Studien konnten zeigen, dass Lächeln beim Sprechen auch ohne visuelle Information wahrnehmbar ist [31, 32]. Äußerungen mit einem mechanischen Lippenspreizen (also ohne emotionale Beteiligung) werden als gelächelter wahrgenommen als solche ohne Spreizung [29]. Wahrnehmbar gelächeltes Sprechen ist auf Veränderungen verschiedener akustischer Parameter zurückzuführen: im Vergleich zu nicht gelächeltem Sprechen ist beim gelächelten Sprechen die Grundfrequenz (F0) erhöht, zum anderen wird bei Vokalen der Frequenzbereich des zweiten Formanten erhöht (durch die Verkürzung des Vokaltrakts an den Lippen und durch Anhebung des Kehlkopfs). Diese Tendenzen haben sich durch spätere Studien mit ganz verschiedenen Arten von Lächeln bestätigt [30, 8].

Nicht in Übereinstimmung mit diesen Befunden ist eine Studie von Caballero et al. [5], bei der Äußerungen in höflicher und in unhöflicher Art aufgenommen wurden. Die akustische Analyse zeigt, dass die höflich produziert Versionen langsamer und bezüglich F0 mit einem niedrigeren globalen Wert und einer Tendenz mit einem Abfall bzw. mit einem geringeren Anstieg bei Ja-Nein-Fragen. Auch wenn es keine allgemein gültige prosodische Gestaltung von Höflichkeit zu geben scheint, so halten die Autoren fest suggest "that prosodic cues routinely and potently interact with other sources of information to allow listeners to generate inferences about im/politeness".

Eine wichtige Unterscheidung sollte gemacht werden zwischen einerseits Lächeln und andererseits Lachen, auch wenn beide Konzepte ähnliche Funktionen teilen (man vergleiche z.B. den Ausdruck "er/sie *lacht* mit mir", wenn diese Personen zu einem *lächelt*). Lachen kann in sehr großer Variabilität und Komplexität auftreten [35], unter anderem als "Sprech-Lachen". Diese Art von Lachen, welches synchron mit Sprechen auftritt, ist in erster Linie durch einen hohen Grad an Behauchtheit und an einer Vibrato-ähnlichen Stimmqualität gekennzeichnet [34] und unterscheidet sich daher von gelächeltem Sprechen auch bezüglich akustischer Parameter, siehe auch die Studie von Erickson et al. [12], in der die sprech-synchronen Formen von Lachen, Lächeln und Weinen akustisch unterschieden werden.

Die Wahrnehmung von Lächeln als sozialem Signal hängt stark davon ab, wie intensiv Lächeln empfunden wird. Die Studie von Émond et al. [11] zeigt, dass Zuhörer länger brauchen, um ein schwaches Lächeln zu erkennen im Vergleich zu einem intensiven Lächeln.

Weitere Untersuchungen sind unbedingt erforderlich, um ein differenzierteres Bild der bekannten und auch der weniger gut untersuchten Parameter zu bekommen, z.B. bezüglich Intensität, Dauer und Stimmqualität. Ein besonderer "Augen- und Ohrenmerk" sollte auf die Wahrnehmung von gelächeltem Sprechen gelegt werden, speziell auch zum Timing von Lächeln in Dialogen. Weitere Desiderate betreffen die empfundene Intensität des Lächelns und die Wahrnehmung von Lächeln über Modalitäten hinweg.

4 Mögliche Applikationen

Angenommen, wir sind in der Lage mit einer Sprachsynthese ein wahrnehmbares Lächeln in synthetisierter Sprache zu erzeugen – in welchen Applikationen könnte dieser Effekt nützlich sein und was sollte man aus der Sicht der Benutzer beachten?

4.1 Sprechprothesen

Eine Sprechprothese ist ein Sprachsynthese-Tool, das einen Ersatz für die eigene Stimme darstellt, z.B. wenn die eigene Stimme krankheitsbedingt verloren gegangen ist (das beste Beispiel hierfür war der Physiker Stephen Hawking). In den vorhergehenden Kapiteln haben wir gesehen, dass es vielfältige Möglichkeiten gibt, Lächeln beim Sprechen, v.a. in Interaktion einzusetzen. Ein gänzlicher Verzicht auf Lächeln (und andere affektive Regungen) erzeugt

den Eindruck eines "Pokerface" oder einen "Buster-Keaton-Effekt". Ein aktiver Benutzer einer Sprechprothese hat wahrscheinlich auch den Wunsch einen Satz bzw. einen Teil davon hörbar lächelnd erzeugen zu lassen. In solchen Fällen würde sich die entsprechende Markierung des Textabschnittes mit einem Markup-Befehl "smiled" anbieten, falls ein solches Feature zur Verfügung stünde. Ferner könnte es möglich sein die Intensität des Lächelns zu kontrollieren, eventuell sogar ob es sich um ein Duchenne-Lächeln handelt oder nicht. Einige vordefinierte Ausdrücke, die routinemäßig benutzt werden, um Höflichkeit und Freundlichkeit zu signalisieren, z.B. bei Gruß- und Dankesworten, könnten in einer bestimmten Variation bereitgestellt werden.

4.2 Audiobücher

Audiobücher ist ein weites Feld der Anwendungen synthetischer Sprache, z.B. in der Blizzard Challenge. Direkte Rede verschiedener Charaktere können in fiktiver Literatur entweder durch verschiedene Schauspieler oder durch denselben Schauspieler mit unterschiedlichen Stimmqualitäten für verschiedene Personen stimmlich porträtiert werden. Eine Voraussetzung für einen angemessen Einsatz gelächelter Sprache wäre hierbei ein Textanalyse-Tool, das relevante Abschnitte mit direkter Rede findet, in der gelächelt werden soll. Dies könnte zum einen durch das Auffinden von Wörtern wie lächeln, grinsen, verschmitzt, freundlich erfolgen oder durch eine "Sentiment analysis", die auf Freundlichkeit, Höflichkeit und weitere Funktionen von Lächeln ausgelegt ist.

4.3 Persona in "Companion Technology"

Im Gegensatz zu virtuellen Agenten, deren sichtbare Dynamik im Gesichtsausdruck mit hoher Qualität animiert werden kann [26], bieten die meisten Köpfe bzw. Gesichter bei sozialen Robotern nicht die Möglichkeit ein sichtbares Lächeln zu erzeugen. Aus diesem Grund könnte es hilfreich sein, ein hörbares Lächeln als soziales Signal zu generieren, falls dabei ein "uncanny valley"-Effekt vermieden werden kann. Zumindest bei virtuellen Agenten wird visuell übertragenes Lächeln als freundlicher und attraktiver empfunden und kann auch zu dem Eindruck einer verstärkten Extrovertiertheit beitragen [6].

Eine besondere Dimension bekommen soziale Roboter im Umgang mit Kindern als Benutzer, z.B. als Tutor bei Krankheitsmanagement oder als Trainingstool für Autisten. Im Allgemeinen scheint in der Interaktion zwischen Kindern mit sozialen Robotern ein erhöhter Grad an Familiarität und Vertrautheit ein wichtiges Thema zu sein, wobei non-verbales Verhalten, Feedback und Interaktionsmanagement zu wichtigen Komponenten bei der Modellierung werden [3]. Lächeln, auch hörbares Lächeln, kann hierbei eine wichtige Rolle spielen.

4.4 Dialogsysteme

Ward & Nakagawa [37] präsentieren einen Ansatz, bei dem Dialogsysteme sich dem Sprechtempo der Benutzer anpassen. Denkbar wäre natürlich auch eine Anpassung des Dialogsystems bezüglich des Lächelns, vorausgesetzt man kann das Lächeln des menschlichen Benutzers verlässlich erkennen. In der anderen Richtung, d.h. bezüglich der Adaption der menschlichen Benutzer an die sprechenden Maschinen, studierten Oviatt et al. [28] Grundschulkinder, die sich mit "embodied conversational agents" (verschiedene Synthesestimmen) unterhielten. Die Mehrheit der Versuchspersonen konvergierten zu den Stimmen der Synthesen bezüglich der Pausenstruktur und der Intensität. Diese Art von Konvergenz konnte auch in einer Studie von Krämer et al. [17] bezüglich des Lächelns beobachtet werden. Die (erwachsenen) Versuchspersonen lächelten länger mit sozialen Robotern, wenn auch die Roboter ein (visuelles) Lächeln zeigten. Auch wenn hier das gegenseitige Lächeln auf Seiten der Benutzer erhöht wurde, so bleibt noch unklar, ob Menschen im Allgemeinen mit Maschinen interagieren

möchten, wenn diese sicht- und/oder hörbar lächeln. Kennedy et al. [15] z.B. zeigten, dass 9-Jährige, die soziale Roboter als Lerntutoren benutzten, eine bessere Leistung erzielten, wenn die Roboter nicht in einer freundlichen Weise interagierten.

5 Gelächelte synthetische Sprache

Eine große Herausforderung besteht darin, das begrenzte Wissen über die Details gelächelten Sprechens für verschiedene Synthesemethoden auszunutzen. Eine weitere Herausforderung ist, Lächeln in Dialogen zu modellieren, da hier zusätzlich zur Signalerzeugung zeitliche, diskursrelevante und auch kulturelle Komponenten mit ins Spiel kommen. Nicht zuletzt muss man sich der Frage stellen, wie die erfolgte Umsetzung der Generierung und Modellierung in einer bestimmten Applikation bewertet werden kann.

5.1 Signalgenerierung

Der Ansatz, der am vielversprechendsten aus der Perspektive der Kontrolle und Generierung zu sein scheint ist die artikulatorische Synthese. Bei dieser Methode lassen sich die akustischen Auswirkungen eines stark verkürzten Vokaltrakts am besten kontrollieren, da zusätzlich zur Kontrolle der Grundfrequenz hier explizit der Kehlkopf angehoben und die Lippen gespreizt werden können. Experimentell konnten z.B. Lasarcyk & Trouvain [20] mit artikulatorischer Synthese zeigen, dass F0 der primäre akustische Faktor ist, der bei der Wahrnehmung des Intensitätsgrads des Gelächelten verantwortlich ist, der am besten bei ungerundeten Vokalen arbeitet.

Auch wenn in artikulatorischer Synthese die sprech-synchronen paralinguistischen Veränderungen der Artikulation abbilden lassen, so ist diese Methode aus einer perzeptuellen Sichtweise die am wenigsten benutzerfreundliche, da die Qualität von einem menschen-ähnlichen am weitesten entfernt ist.

Konkatenative Synthese hat den Vorteil, dass akustische Abschnitte natürlicher Sprache verarbeitet werden. Der Nachteil besteht aber darin, dass die gewünschten Effekte entweder bereits in der akustischen Datenbank vorliegen muss oder die erzeugten Signalketten nachträglich in den gewünschten Parametern verändert werden müssen. Eine gesamte Datenbank zusätzlich in einer gelächelten Version zur Verfügung zu stellen ist unrealistisch. Die Manipulation von F0 hingegen ist wenig problematisch, die nachträgliche Änderung von F2 hingegen schon, obwohl Arias et al. [1] in audio-visuellen Experimenten die ersten beiden Formanten erfolgreich manipulieren konnten.

Aktuellere Ansätze, wie Unit-Selection, Hidden-Markov-Modelle (HMM) oder Neuronale Netze (NN), basieren auf größeren annotierten Datenbanken eines Sprechers. Es erscheint jedoch unrealistisch, gelächelte Sprache in der Größenordnung von Stunden zu erheben. Damit wäre die Nutzung von Unit-Selection auf wenige, hochfrequente Wörter und Ausdrücke beschränkt, falls nicht auf nachträgliche Signalmanipulation zurückgegriffen werden soll. Für parametrische Ansätze (HMM, NN) könnten Methoden des Sprechermorphings ausgelotet werden, die auf Basis weniger Daten gelächelter Sprache eine Anpassung der akustischen Repräsentation umsetzen, indem beispielsweise der Parameterraum des verwendeten Sprachkodierers anhand weniger Stützparameter transformiert wird, oder Transfer-Learning-Ansätze genutzt werden [38, 7, 2, 25]. El Haddad et al. [10] berichten von ersten Ansätzen zur Modellierung gelächelten Sprechens mit HMM-basierten Synthese, die aber nur einen Anfang darstellen kann.

5.2 Anwendung und Modellierung

Für die Nutzung gelächelter synthetischer Sprache in echten Anwendungen bedarf es einer kontextuell angemessenen Ansteuerung der Synthese. Dies betrifft selbst beim reinen Vorlesen, z.B. Audiobüchern, der kulturell und inhaltlich angemessenen Auswahl gelächelter Abschnitte und Dauern. Für eine automatische symbolische Annotation der zu lächelnden Abschnitte bedarf es also vermutlich neben einem sprach- bzw. kulturspezifischen Modell auch einer Analyse der Semantik und des Sentiments des vorliegenden Textes.

Für interaktive Anwendungen müssten zudem diskursabhängige soziale Signale generiert werden, falls sie als gelächelte Sprache umgesetzt werden können, z.B. reziprokes Lächeln. Dafür muss jedoch auch geklärt werden, wie sich Lächeln in menschlicher Kommunikation zwischen audio-visuellen und rein akustischen Situationen unterscheidet.

Eine Evaluierung sollte demnach dem Entwicklungsstand angepasst werden: Erkennbarkeit und Natürlichkeit der Sprachgenerierung als Qualitätsmerkmale müssen genauso überprüft werden wie die Angemessenheit (Natürlichkeit) der Position und Dauer gelächelter Sprache. Aufgrund der vielfachen funktionalen Verwendungen von Lächeln muss also auch Intention und Interpretation verglichen werden.

6 Diskussion

Eingangs haben wir drei Fragen gestellt, die in dieser Sektion diskutiert werden sollen. Soll man bei einer synthetischen Stimme erkennen können, dass sie lächelt? Wir denken, dass es etliche Anwendungen gibt, in denen diese Erkennung sinnvoll und nützlich ist, da sie eine bessere Interpretation des Gesagten erlaubt. Für Sprechprothesen und Audiobücher ist der Nutzen unmittelbar einsichtig. Für dialogische Applikationen, in denen diese spezielle Sprechkompetenz den Benutzern klar gemacht worden ist, kann dieses Extra sich zu einem Vorteil entwickeln. Es kann aber auch davon ausgegangen werden, dass durch diese ungewohnte und unerwartete Modifikation der Sprechweise auf Seiten der Benutzer zunächst Irritationen entstehen können. Die erste Frage ist eng verknüpft mit der zweiten: Sollen sprechfähige Maschinen in Form von Computern, virtuellen Agenten oder sozialen Robotern lächeln können? Die Anthropomorphisierung nicht-menschlicher Objekte würde dadurch verstärkt werden, was sich positiv nutzen lässt in Situationen der Mensch-Maschine-Interaktion, in denen es vorteilhaft ist, wenn mehr Vertrauen aufgebaut werden kann. Es könnte aber auch zu enttäuschenden Erwartungen bezüglich der sozialen Kompetenz führen oder gar zu einem ein "uncanny valley of mind" [14] führen, was sicherlich nicht wünschenswert ist.

Wie sollte ein solches Vorhaben angegangen werden? Zwar gibt es für die Generierung gelächelter synthetischer Sprache den ein oder anderen Ansatz, der lohnt weiter verfolgt zu werden, von einem überzeugenden Umsetzen der Aufgabe sind wir aber (sehr) weit entfernt. Dies gilt genauso für die Modellierung und die Evaluierung. Ein Grund für diese Lage liegt in dem noch weitgehend fehlenden Verständnis der Komplexität des Themas.

Forschung zu Lächeln beim Menschen konzentriert sich weitestgehend auf den visuellen Kanal. Die Hauptuntersuchungsobjekte sind oftmals statische Fotos von Gesichtern (häufig ohne Brille und Bart). Die zeitliche Dynamik fehlt in aller Regel, genauso die wechselnde Intensität und der situative (und auch verbalisierte) Kontext. Dies sind alles sehr wichtige Komponenten für die Modellierung von Lächeln in gesprochener Sprache (sowohl für natürliche als auch für synthetisierte).

Obwohl unsere Überlegungen auf reine auditive Aspekte synthetischer Sprache abzielen, so sollten natürlich auch audio-visuelle Aspekte von Lächeln in der Sprachsynthese mitbedacht werden, z.B. zum Einsatz in "embodied conversational agents" und in sozialen Robotern. Dabei wäre es enorm wichtig, wenn die Grundlagenforschung mehr Klarheit erbringen würde

bezüglich 1) wann genau, 2) zu welchem Grad und 3) für welchen Zweck bzw. in welcher Funktion Menschen in gesprochener Interaktion lächeln. Es gibt auch eine große Forschungslücke bezüglich der phonetischen Aspekte gelächelter Sprache. Wie hängen hierbei Akustik und Wahrnehmung zusammen und wie Akustik und Sprachproduktion? Wie verhalten sich die visuellen Informationen zu den akustischen, insbesondere in Dialogen, sowohl in der Mensch-Mensch als auch in der Mensch-Maschine-Interaktion? Wir sehen daher in den dargelegten Überlegungen zu hörbarem Lächeln einen weiteren kleinen Baustein in der Weiterentwicklung des "social signal processing" [36].

Literatur

[1] Arias, P., Belin, P. & Aucouturier, J.-J. 2018. Auditory smiles trigger unconscious facial imitation. Current Biology 28, paper R782.

[2] Arik, S.O., Chen, J., Peng, K., Ping, W. & Zhou, Y. 2018. Neural voice cloning with a few samples. *arXiv preprint* arXiv:1802.06006.

[3] Belpaeme, T. et al. 2018. Guidelines for designing social robots as second language tutors. *International Journal of Social Robotics* 10, pp. 325-341.

[4] Burgoon, J., Buller, D., Hale, J., & Turck, M. 2018. Relational messages associated with nonverbal behaviors. Human Communication Research 10, pp. 351–378.

[5] Caballero, J.A., Vergis, N., Jiang, X. & Pell, M.D. 2018. The sound of im/politeness. *Speech Communication* 102, pp. 39-53.

[6] Cafaro, A., Vilhjálmsson, H.H., Bickmore, T., Heylen, D., Jóhannsdóttir, K.R. & Valgarðsson, G.S. 2012. First impressions: users' judgments of virtual agents' personality and interpersonal attitude in first encounters. Proc. *12th Int'l Conf. Intell. Virtual Agents,* 1-14.

[7] Doddipatla, R., Braunschweiler, N. & Maia, R. 2017. Speaker adaptation in DNN-based speech synthesis using d-vectors. Proc. *Interspeech*, Stockholm, pp. 3404–3408.

[8] Drahota, A., Costall, A. & Reddy, V. 2008. The vocal communication of different kinds of smile. *Speech Communication* 50(4), pp. 278-287.

[9] Ekman, P. & Friesen, W.V. 1982. Felt, false, and miserable smiles. *Journal of Nonverbal Behavior* 6(4), pp. 238-258.

[10] El Haddad, K., Çakmak, H., Moinet, A., Dupont, S. & Dutoit, T. 2015. An HMM approach for synthesizing amused speech with a controllable intensity of smile. *IEEE International Symposium on Signal Processing and Information Technology*, pp. 7-11.

[11] Émond, C., Rilliard, A. & Trouvain, J. 2016. Perception of smiling in speech in different modalities by native vs. non-native speakers. Proc. *Speech Prosody*, Boston, pp. 639-643.

[12] Erickson, D., Menezes, C. & Sakakibara, K. 2009. Are you laughing, smiling or crying? Proc. *APSIPA Summit and Conference*, pp. 529-537.

[13] Floyd, K. & Erbert, L. 2003. Relational message interpretations of nonverbal matching behavior: an application of the social meaning model. *Journal of Social Psychology* 143, pp. 581–597.

[14] Gray, K. & Wegner, D.M. 2012. Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition* 125, pp. 125-130.

[15] Kennedy, J., Baxter, P. & Belpaeme, T. 2017. The impact of robot tutor nonverbal social behavior on child learning. Frontiers in ICT Human-Media Interaction 4, article 6.

[16] Keltner, D. 1995. Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology* 68, pp. 441–454.

[17] Krämer, N., Kopp, S., Becker-Asano, C. & Sommer, N. 2013. Smile and the world will smile with you — The effects of a virtual agent's smile on users' evaluation and behavior. *International Journal of Human-Computer Studies* 71(3), pp. 335-349.

[18] Krumhuber, E., Manstead, A.S.R., Cosker, D., Marshall, D., Rosin, P.L. & Kappas, A. 2007. Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emo-*

tion 7, pp. 730–735.
[19] Krys et al. 2016. Be careful where you smile: Culture shapes judgments of intelligence and honesty of smiling individuals. *Journal of Nonverbal Behavior* 40, pp. 101–116.
[20] Lasarcyk, E. & Trouvain, J. 2008. Spread lips + raised larynx + higher F0 = smiled speech? - An articulatory synthesis approach. Proc. 8th *International Speech Production Seminar* (ISSP '08), Strasbourg, pp. 345-348.
[21] Lau, S. 1982. The effect of smiling on person perception. *Journal of Social Psychology* 117, pp. 63–67.
[22] Mehu, M., Little, A.C. & Dunbar, R.I. 2008. Sex differences in the effect of smiling on social judgments: an evolutionary approach. *Journal of Social, Evolutionary, and Cultural Psychology* 2, pp. 103–121.
[23] Moore, M.M. 1985. Non-verbal courtship patterns in women: context and consequences. *Ethology and Sociobiology* 6, pp. 237-247.
[24] Mui, P.H.C., Goudbeek, M.B., Swerts, M.G.J. & Hovasapian, A. 2017. Children's nonverbal displays of winning and losing: Effects of social and cultural contexts on smiles. *Journal of Nonverbal Behavior* 41, pp. 67-82.
[25] Nachmani, E., Polyak, A., Taigman, Y. & Wolf, L. 2018. Fitting new speakers based on a short untranscribed sample. Proc. *Int'l Conf. on Machine Learning*, pp. 3683-3691.
[26] Ochs, M., Niewiadomski, R. & Pelachaud, C. 2010. How a virtual agent should smile? Morphological and dynamic characteristics of virtual agent's smiles. Proc. *Int'l Conf. on Intelligent Virtual Agents*, pp. 427-440.
[27] Okubo, M., Ishikawa, K., Kobayashi, A., Laeng, B., & Tommasi, L. 2015. Cool guys and warm husbands: The effect of smiling on male facial attractiveness for short- and long-term relationships. *Evolutionary Psychology* 13(3), pp. 1-8.
[28] Oviatt, S.L. Darves, C. & Coulston, R. 2004. Toward adaptive conversational interfaces: Modeling speech convergence with animated personas. *ACM Transactions in Computer-Human-Interaction* 11(3), pp. 300-328.
[29] Robson, J. & Beck, J. M. 1999. Hearing smiles – Perceptual, acoustic and production aspects of labial spreading. Proc. *14th Int'l Congress of Phonetic Sciences* (ICPhS), San Francisco, pp. 219-222.
[30] Schröder, M., Aubergé, V. & Cathiard, M.-A. 1998. Can we hear smiles? Proc. *Conference on Spoken Language Processing* (ICSLP), vol. 3, pp. 559-562.
[31] Tartter, V. C. 1980. Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception and Psychophysics* 27(1), pp. 24-27.
[32] Tartter, V.C. & Braun, D. 1994. Hearing smiles and frowns in normal and whisper registers. *Journal of the Acoustical Society of America* 96(4), pp. 2101-2107.
[33] Thibault, P., Levesque, M., Gosselin, P. & Hess, U. 2012. The Duchenne marker is not a universal signal of smile authenticty – but it can be learned! *Social Psychol.* 43, 215-221.
[34] Trouvain, J. 2001. Phonetic aspects of "speech-laughs". Proc. *Conf. Orality & Gestuality* (ORAGE) 2001, Aix-en-Provence, pp. 634-639.
[35] Truong, K.P., Trouvain, J. & Jansen, M.-P. 2019. Towards an annotation scheme for complex laughter in speech corpora. Proc. *Interspeech*, Graz, pp. 529-533.
[36] Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F. & Schröder, M. 2012. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing* 3(1), pp. 69–87.
[37] Ward, N. & Nakagawa, S. 2004. Automatic user-adaptive speaking rate selection. *International Journal of Speech Technology* 7, pp. 259-268.
[38] Wu, Z., Swietojanski, P., Veaux, C., Renals, S. & King, S. 2015. A study of speaker adaptation for DNN-based speech synthesis. Proc. *Interspeech*, Dresden, pp. 879–883.