MASCHINELLE KLASSIFIKATION VON ARTIKULATIONSBEWEGUNGEN IM RAHMEN EINER VISUELLEN ARTIKULATIONSSCHULUNG FÜR GEHÖRLOSE UND SCHWERHÖRIGER KINDER

Ralf Winkler, [1,2] *Gunter Uhlmann* [2] *und Gerd Schneider* [2]

[1] *Kommunikationswissenschaft, Technische Universität Berlin*
[2] *Rehabilitationstechnik/ Neue Medien, Humboldt-Universität zu Berlin*
ralf.winkler@tu-berlin.de

Kurzfassung: Unterschiedliche technische Hilfsmittel wurden entwickelt, um gehörlose und profund schwerhörige Kindern beim Erlernen der Lautsprache zu unterstützen. Die Systeme sind in der Lage, selbstständig, ohne Eingriff eines Therapeuten, den Erfolg einer Äußerung zu bewerten. Probleme bereitet dagegen allen Systemen die Erzeugung von Hinweisen zur Korrektur im Falle von Misserfolg.
Es wird untersucht, in welchem Maße Bewegungen der äußeren Artikulatoren zur Erzeugung von Hinweisen zur Korrektur von Sprechbewegungen geeignet sind. Ein Schema für eine explizite Repräsentation korrekter Artikulation wird vorgeschlagen und anhand der Vokale des Deutschen evaluiert. Anhand einer großen Anzahl von Bewegungsdaten eines Sprechers sowie von insgesamt 14 Sprechern wird getestet, in welchem Maße diese explizite Repräsentation geeignet ist, zwischen korrekten und defizitären Artikulationen zu unterscheiden.
Die Ergebnisse der Analysen der Bewegungsdaten eines einzelnen Sprechers zeigen, dass die explizite Repräsentation nahezu fehlerfrei sechs Vokalgruppen abbildet. Hieraus lässt sich der Schluss ziehen, dass die hier definierten artikulatorischen Merkmale grundsätzlich zur Spezifikation der verwendeten Vokalgruppen geeignet sind. Die Klassifikation der Bewegungsdaten von zusammen 14 Sprechern ist trotz auftretender Verwechslungen ebenfalls gut. Dieses Ergebnis zeigt, dass eine sprecher-unabhängige Repräsentation von Artikulationen nach dem hier vorgeschlagenen Schema möglich ist. Da das Ziel der trainierenden Kinder das Erzeugen verständliche Sprache ist, sollte die Bewertung des Erfolges weiterhin primär auf der Ebene des Audiosignals erfolgen. Das hier vorgeschlagene Schema könnte von einem technischen System verwendet werden, bei Defiziten durch Vergleich zwischen Sprechbewegung und hier vorgeschlagener Repräsentation selbstständig Hinweise zur Korrektur der Sprechbewegung an den Lerner zu generieren.

1 Einleitung

Normal hörende Kinder optimieren im Rahmen des Spracherwerbs ihre Aussprache durch ständigen unbewussten Abgleich der eigenen Sprachlaute mit denen ihrer Interaktionspartner. Eine notwendige Bedingung für die auditive Rückmeldung ist eine normale Hörfähigkeit. Darum ist gehörlosen bzw. profund schwerhörigen Kindern der Erwerb der Lautsprache auf normalem Wege nicht möglich.
Gehörlose Kinder können Lautsprache nur explizit, normalerweise im Rahmen sprachtherapeutischen Unterrichts, erlernen, wo die eigene auditive Rückmeldung durch die Rückmeldung des Lehrers ersetzt wird. Um, zusätzlich zum Unterricht, gehörlose Kinder beim Erlernen von Lautsprache zu unterstützen, wurden z.B. technische Systeme entwickelt [1, 11, 12, 13, 10], welche

das Sprachsignal des Kindes aufzeichnen, analysieren, mit prototypischen Äußerungen vergleichen und eine Rückmeldung über den Erfolg generieren [9]. In diesen Fällen wird die fehlende auditive Rückmeldung des Lernenden durch die Rückmeldung eines technischen Systems übernommen. Am Beispiel von Kindern mit Cochlea-Implantat wurde ein effizienter Einsatz derartiger technischer Hilfsmittel in der Sprachtherapie (*engl. computer-based speech training, CBST*) bereits nachgewiesen [3]. Da der Lernvorgang ohne Mitwirkung eines Therapeuten bei den genannten Systemen jedoch nach dem trial-and-error-Prinzip abläuft, sind derartige Systeme weniger für eigenständiges Üben als vielmehr für das Trainieren der Aussprache z.B. im Rahmen des Fremdspracherwerbs (*engl. computer-assisted pronunciation training, CAPT*) geeignet.
Technische Systeme, welche den Kindern eigenständiges Üben ermöglichen, müssen in der Lage sein, im Falle defizitärer Artikulationen über die Bewertung des Erfolges hinaus aussagekräftige Korrekturvorschläge zu erzeugen. Ein solches System muss zur Erzeugung eines Korrekturhinweises entweder artikulatorische Merkmale aus dem Audiosignal rekonstruieren (*acoustic articulatory inversion*) oder direkt aufzeichnen und auf eine explizite Repräsentation (Konzept) von korrekter bzw. defizitärer Artikulation zugreifen können.
Nach dem erstgenannten Prinzip arbeitet das System ARTHUR [2], welches das Audiosignal zusammen mit visuellen Merkmalen zur Rekonstruktion der Bewegung der Artikulatoren heranzieht [6]. Die Modellierung der Inversion erfolgt jedoch auf der Basis von Sprach- und Bewegungsdaten eines einzelnen Sprechers. Somit enthält das erstellte Modell der Inversion alle sprecherspezifischen Eigenheiten der Sprechbewegung sowie seine morphologischen Randbedingungen. Die Anpassung des Modells auf Daten beliebiger Sprecher erfordert eine komplizierte Anpassung des Audiosignals, wobei bisher unklar ist, mit welcher Genauigkeit die Extraktion der notwendigen Daten möglich ist.
In dieser Arbeit wird ein Schema zur Erzeugung einer aussagekräftigen Rückmeldung vorgeschlagen, welches nach dem zweiten Prinzip arbeitet. Sprachsignale entstehen durch die koordinierte Bewegung innerer und äußerer Artikulatoren, die physiologisch miteinander verbunden sind [4] und deren Bewegungen nicht willkürlich voneinander stattfinden. Die Bewegung zumindest der äußeren Artikulatoren, wie z.B. der Lippen und des Unterkiefers, können durch technische Verfahren direkt gemessen werden. Aus den artikulatorischen Daten einer Vielzahl von Sprechern wird ein Konzept über korrekte bzw. defizitäre Artikulationsbewegungen abgeleitet. Das vorgeschlagene Schema arbeitet ohne Bezug zum akustischen Signal. Sprechbewegungen werden durch Vergleich mit dem Konzept korrekter Artikulation bezüglich des Erfolges bzw. Misserfolges bewertet. Bei Misserfolg werden durch Vergleich zwischen defizitärer Sprechbewegung und dem Konzept konkrete Hinweise zur Korrektur der Sprechbewegung generiert.

2 Daten & Methoden

2.1 Korpus & Aufnahmen

Das Sprachmaterial besteht aus Aufnahmen von insgesamt 15, vorwiegend weiblichen Sprechern im Alter zwischen 20 und 36 Jahren. Es wurden insgesamt 15 Vokale des Deutschen analysiert. Jeder Vokal ist in einem Trägerwort in der ersten Silbe zwischen zwei labialen Verschlüssen eingebettet (s. Tab. 1).
Das System zur Aufnahme der Bewegungsdaten besteht aus fünf Infrarot-Kameras (VICON M-Cam) mit Infrarot-Diodenkranz (*motion capture system* VICON 512). Für die Etikettierung der Marker und zur Rekonstruktion ihrer Position wird die Software VICON Workstation (Version 4.6) verwendet. Es wurden ferner Audiosignale und digitale Bilddaten (DV Kamera) angefertigt, die während der Analyse der geometrischen Daten zur Interpretation der Bewegungen der

Tabelle 1 - Wörter mit den zu analysierenden Vokalen, analysierte Vokale befinden sich in der ersten Silbe (Fettdruck).

p**a**pa (Papa)	b**ɪ**məl (Bimmel)	b**ʊ**məln (bummeln)	b**œ**tçɐ (Böttcher)
b**aː**biː (Barbie)	b**iː**bɐ (Bieber)	b**uː**bə (Bube)	b**øː**mə (Böhme)
b**ɜ**mə (Bemme)	b**ɔ**məl (Bommel)	-	p**ʏ**pçən (Püppchen)
m**eː**mo (Memo)	p**oː**plɪg (poplig)	m**æː**dəl (Mädel)	b**yː**bçən (Bübchen)

Artikulatoren herangezogen wurden. Die hier vorgestellten Ergebnisse basieren jedoch ausschließlich auf den mittels *motion capturing* erfassten Bewegungsdaten.

2.2 Spezifikation der Vokalgruppen

Die Grenzen zwischen einzelnen Vokalen, z.B. zwischen halb-offenen und offenen, können aus den artikulatorischen Daten nicht beliebig genau bestimmt werden. Andere artikulatorische Dimensionen, z.B. *vorn - zentral - hinten*, werden durch die Bewegung der äußeren Artikulatoren gar nicht abgebildet. Daher wurden die Vokale im Vorhinein derart gruppiert, dass sich die Gruppen jeweils bzgl. eines aus den Bewegungsdaten bestimmbaren Merkmals maximal unterscheiden. Die Bildung von Gruppen (s. Tab. 2) basiert auf den Eigenschaften von Vokalen in [5].

Tabelle 2 - Vokalgruppen mit den intendierten Ausprägungen der artikulatorischen Merkmale.

Vokalgruppe	Kieferöffnung	Lippenrundung	Gespanntheit
[a,E] | *[a:,E:]*	offen	ungerundet	ungespannt | gespannt
[I] | *[i:,e:]*	geschlossen	ungerundet	ungespannt | gespannt
[O,U,9,Y] | *[o:,u:,2:,y:]*	.	gerundet	ungespannt | gespannt

2.3 Artikulatorische Merkmale

Jede Ziel-Artikulation wird bezüglich der folgenden drei artikulatorischen Merkmale bewertet: Kieferöffnung: *offen* (z.B. aː in Barbie) vs. *geschlossen* (z.B. eː in Memo); Lippenrundung: *gerundet* (z.B. øː in Böhme) vs. *ungerundet* (z.B. eː in Memo) sowie Gespanntheit: *gespannt* (z.B. iː in Miete) vs. *ungespannt* (z.B. ɪ in Mitte).

Kieferöffnung: Das Merkmal *Kieferöffnung* wurde durch die euklidische Distanz zwischen einem Marker am Kinn und einem Marker am Nasenrücken parametrisiert, welche auf der medio-sagittalen Ebene des Gesichtes appliziert wurden. Der Marker auf dem Nasenrücken wurde so angebracht, dass seine relative Position beim Sprechen weitestgehend konstant bleibt. Von der tatsächlichen Distanz wird die zuvor bestimmte Distanz dieser Marker in Ruheposition des Mundes subtrahiert, um unabhängig von morphologischen Merkmalen des Gesichtes artikulatorische Bewegungen zu analysieren. Der Zeitpunkt der maximalen Kieferöffnung wird als Zeitpunkt der Erreichung der artikulatorischen Zielposition interpretiert. Der Wert der Distanz-Kurve an diesem Zeitpunkt wird für das Merkmal *Kieferöffnung* herangezogen.

Lippenrundung: Ziel der Lippenrundung ist die aktive Verlängerung des Sprechtraktes im Bereich der Lippen, um die ersten zwei Formanten abzusenken. Zur Parametrisierung wird

die Korrelation zwischen dem Vorstülpen der Lippen und der damit einher gehenden Verkleinerung der euklidischen Distanz zwischen linkem und rechtem Mundwinkel ausgenutzt. Um morphologische Eigenschaften wie die Länge der Lippen aus den Daten zu entfernen, werden nur Bewegungen der Mundwinkel analysiert, indem der zuvor bestimmte Abstand zwischen den Mundwinkeln in Ruheposition des Mundes abgezogen wird. Maßgeblich für den Wert des Merkmals *Lippenrundung* ist die Amplitude der Stülp- bzw. Spreiz-Bewegung am Zeitpunkt der Erreichung der artikulatorischen Zielposition.

Gespanntheit: Die Gespanntheit schlägt sich im Deutschen in der Dauer des Vokals nieder. In den Bewegungsdaten wird die Vokaldauer durch die Dauer der Kieferbewegung approximiert. Diese wird in der ersten Ableitung der Distanzkurve (s. Merkmal *Kieferöffnung*), welche zur Kurve des Verlaufs der Geschwindigkeit führt, bestimmt. Der Beginn der Kieferöffnung wird am Zeitpunkt der maximalen Geschwindigkeit der Öffnung des Unterkiefers, das Ende am Zeitpunkt der maximalen Geschwindigkeit der Schließbewegung, festgelegt. Das Merkmal *Gespanntheit* wird als zeitlicher Abstand zwischen jeweiligem Anfangs- und Endpunkt definiert.

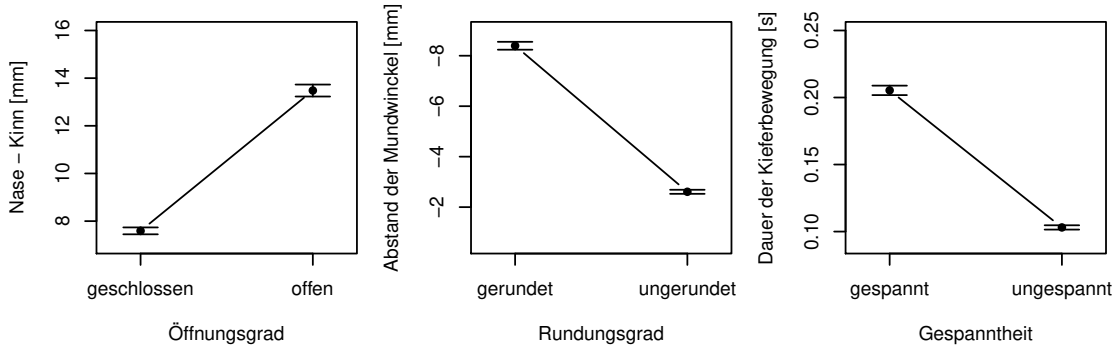
2.4 Klassifikation & Rückmeldung

In der ersten Schicht wird eine Sprechbewegung bezüglich der drei artikulatorischen Merkmale bewertet. Für jedes Merkmal einzeln wurde ein Entscheidungsbaum (C4.5 in [8]) mit allen zur Verfügung stehenden Daten generiert. Ein Klassifikator bewertet eine Äußerung bezüglich der Kieferöffnung, ein zweiter bezüglich der Lippenrundung und ein dritter bezüglich der Gespanntheit. Jeder dieser Klassifikatoren gibt als Ergebnis ein binäres Attribut aus, welches wiederum einen Eingang zu einem weiteren Entscheidungsbaum (zweite Schicht, repräsentiert durch Tab. 2) darstellt. Wurde in der zweiten Schicht eine zu bewertende Sprechbewegung der intendierten Vokalgruppe zugeordnet, haben die drei Klassifikatoren der ersten Schicht korrekt klassifiziert. Folglich war die Sprechbewegung korrekt. Andernfalls weist die Artikulation Defizite auf und eine Rückmeldung wird generiert.
Zur Erzeugung einer aussagekräftigen Rückmeldung wird zunächst das Defizit identifiziert. Zu diesem Zweck wird das Merkmal (bzw. werden die Merkmale) ermittelt, bei denen auf Grund einer Differenz zwischen intendierter und tatsächlicher Ausprägung in der ersten Schicht falsch klassifiziert wurde. Basierend auf der Diskrepanz zwischen Schwellwert und tatsächlicher Ausprägung kann dann ein aussagekräftiger Hinweis zur Korrektur der Sprechbewegung erzeugt werden.

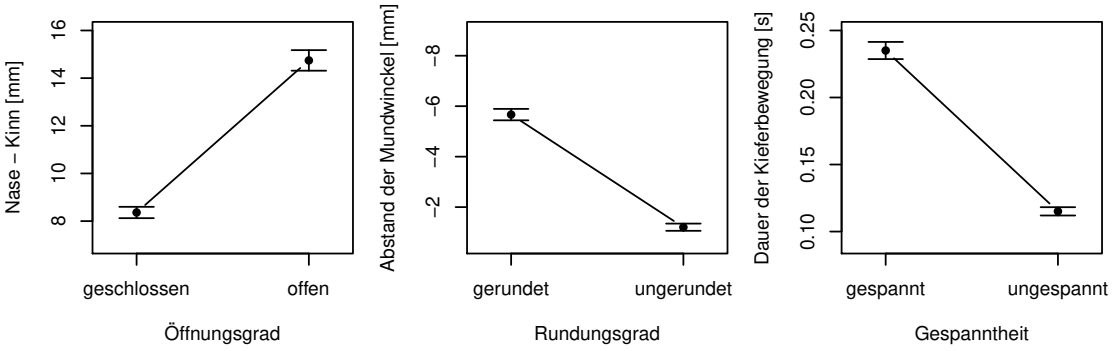
3 Ergebnisse

3.1 Experiment I

Die Analyse der verwendeten artikulatorischen Merkmale gruppiert nach ihren intendierten Ausprägungen für einen Sprecher kann in Abbildung 1(a) eingesehen werden. Bei allen drei Merkmalen zeigen sich ausgeprägte Unterschiede bezüglich der beabsichtigten Ausprägung. Obwohl, relativ zur Neutralstellung des Unterkiefers, bei jeder Silbe eine Kieferöffnung realisiert wird, öffnet sich der Kiefer bei den offenen Vokalen deutlich weiter ($\bar{x}$ = 13,5 mm, SD = 1,7 mm) als bei den geschlossenen Vokalen ($\bar{x}$ = 7,6 mm, SD = 1,6 mm). Abbildung 1(a) (Mitte) zeigt, dass die Bewegung der Mundwinkel bei den ungerundeten Vokalen ($\bar{x}$ = -2,6 mm, SD = 0,7 mm) im Vergleich zu den gerundeten Vokalen ($\bar{x}$ = -8,4 mm, SD = 1,5 mm) deutlich geringer ist. Schließlich zeigen die Analysen (s. Abb. 1(a) (rechts)), dass auch die Dauer der Kieferbewegung, hier als Approximation der Gespanntheit, erheblich zwischen gespannten (($\bar{x}$ = 205 ms, SD = 34,6 ms)) und ungespannten Vokalen ($\bar{x}$ = 103 ms, SD = 14,7 ms) differiert.



(a) ein Sprecher, 12 Wiederholungen, N=180



(b) 14 Sprecher, keine Wiederholungen, N=210

Abbildung 1 - Darstellung der Kieferöffnung, der Lippenrundung sowie der Gespanntheit für die Vokalgruppen, Fehlerbalken zeigen ± 1 SE.

Die Dauer der Kieferbewegung dieses Sprechers ist bei den gespannten Vokalen etwa doppelt so lang im Vergleich zu den ungespannten Vokalen. Die drei Differenzen sind statistisch signifikant (t-tests, zwei-seitig, $p < 0.001$).

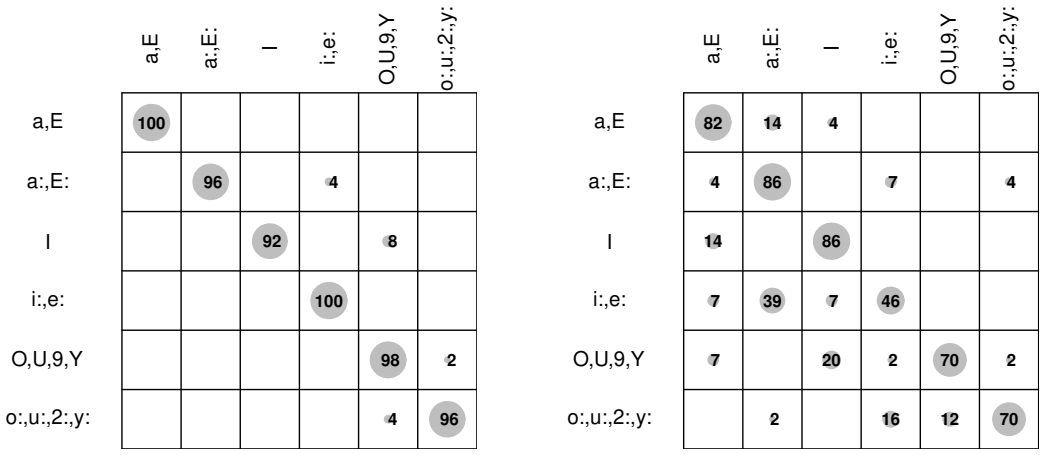
Die festgestellten Differenzen legen nahe, dass eine Klassifizierung von Sprechbewegungen bezüglich der verwendeten Merkmale möglich ist. Tabelle 3 (Exp. I) zeigt die geschätzten Klas-

Tabelle 3 - Kennwerte der drei einzelnen Klassifikatoren, ein Sprecher, zwölf Wiederholungen, N=180 (Exp. I) sowie 14 Sprecher, N=210 (Exp. II), Klassifikationsrate wurde mit 10-facher Kreuzvalidierung approximiert.

Klassifikator	**Klassifikationsrate** [%]		**Trennwert**	
	Exp. I	Exp. II	Exp. I	Exp. II
1 (Kieferöffnung)	98,33	81,43	10,8 mm	10,4 mm
2 (Lippenrundung)	98,89	84,29	-4,5 mm	-4,0 mm
3 (Gespanntheit)	97,22	90,95	131 ms	163 ms

sifikationsraten der einzelnen binären Klassifikatoren der ersten Schicht. Die Klassifikationsraten sind mit Werten zwischen 97% und 99% ausgezeichnet. Nahezu alle Sprechbewegungen werden der intendierten Ausprägung des Merkmals zugeordnet. Ebenfalls angegeben (Tab. 3, Exp. I) sind die durch einmalige Optimierung (C 4.5) ermittelten Schwellwerte zur Bestimmung der prototypischen Ausprägung des jeweiligen Merkmals. Wird für eine Artikulation z.B. eine

Kieferöffnung $\leq$ 10,8 mm gemessen, gibt der Klassifikator 1 (Kieferöffnung) die Klasse *geschlossen*, andernfalls die Klasse *offen*, aus. Basierend auf den Ergebnissen der einzelnen Klassifikatoren der ersten Schicht wird in der zweiten Schicht eine Vokalgruppe vorhergesagt. Der Vergleich zwischen intendierter (Referenzklasse) und klassifizierter Vokalgruppe (Vorhersage) führt zur Verwechslungsmatrix in Abbildung 2(a). Die Vokale der Gruppe *a,E* sowie der Gruppe



(a) Klassifikationsrate: $\bar{x}$ = 96,88%

(b) Klassifikationsrate: $\bar{x}$ = 73,21%

Abbildung 2 - (Normierte) Verwechslungsmatrix des Klassifikationssystems für die Daten eines Sprechers mit zwölf Wiederholungen (a) sowie für 14 Sprecher ohne Wiederholungen (b), Zeilen entsprechen Referenzklassen, Spalten entsprechen der Vorhersage, Zeilensumme jeweils 100%.

i:,e: werden durch das zweistufige Schema ausnahmslos richtig zugeordnet. Defizitäre Kieferöffnung führt zu minimalen Verwechselungen zwischen den Gruppen *a:,E:* und *i:,e:*. Fälschlicherweise der Gruppe *O,U,9,Y* zugeordnete ɪ-Artikulationen resultieren aus Defiziten bei der Formung der Lippenrundung. Ferner sind minimale Verwechslungen zwischen den Gruppen *o:,u:,2:,y:* und *O,U,9,Y* zu beobachten, welche auf Artikulationen von zu langen ungespannten bzw. zu kurzen gespannten Vokalen zurückzuführen sind. Die einzelnen Klassifikationsraten auf der Matrix-Diagonalen ($\geq$ 92%) zeigen jedoch an, dass die beteiligten Merkmale ausgezeichnet geeignet sind, korrekte artikulatorische Ziele zumindest eines Sprechers zu repräsentieren.

3.2 Experiment II

Die Messungen des Merkmals Kieferöffnung (*offen*: $\bar{x}$ = 14,7 mm, SD = 3,2 mm; *geschlossen*: $\bar{x}$ = 8,4 mm, SD = 3,0 mm), der Lippenrundung (*gerundet*: $\bar{x}$ = -5,7 mm, SD = 2,4 mm; *ungerundet*: $\bar{x}$ = -1,2 mm, SD = 1,4 mm) sowie der Gespanntheit (*gespannt*: $\bar{x}$ = 235 ms, SD = 67,8 ms; *ungespannt*: $\bar{x}$ = 115,1 ms, SD = 30,7 ms) legen nahe (vgl. Abb. 1(b)), dass die jeweiligen Abstände auch bei Mittelung über 14 Sprecher hinreichend groß sind, um basierend auf einem Schwellwert zwischen den jeweiligen Ausprägungen der drei artikulatorischen Merkmale zu unterscheiden. Die Differenzen zwischen den Ausprägungen aller drei Merkmale sind statistisch signifikant (t-tests, zwei-seitig, $p < 0.001$).

Die drei Klassifikationsraten der ersten Schicht liegen oberhalb von 80% (vgl. Tab. 3 (Exp. II)). Die beste Erkennung wird mit durchschnittlich 90,95% bei der Klassifikation bezüglich der Gespanntheit erreicht. Ebenfalls zuverlässig können die Daten bezüglich der Lippenrundung (84,29%) sowie bezüglich der Öffnung des Unterkiefers (81,43%) klassifiziert werden.

Die resultierende Verwechslungsmatrix des Klassifikators der zweiten Schicht (s. Abb. 2(b)) zeigt, dass drei Vokalgruppen, *a,E* und *a:,E:* sowie *I*, relativ stabil mit einer Wahrscheinlichkeit oberhalb von 80% korrekt klassifiziert werden. Nahe an dieser Grenze liegt mit jeweils 70% die Klassifikationsrate der Gruppe *U,Y,O,9* und *o:,u:,2:,y:*. Auffallend schwach ausgeprägt ist die Erkennung von Sprechbewegungen der Gruppe *i:,e:* (46%), wo hauptsächlich Verwechslungen mit der Gruppe *a:,E:* auftreten. Diese hängen wahrscheinlich mit der Übungssituation der Sprecher zusammen. Die betroffenen Vokalgruppen unterscheiden sich bezüglich der intendierten Öffnung des Unterkiefers. Während der Aufnahmen ist bei vielen Sprechern eine ausgeprägte Kieferbewegung festzustellen, was bei den geschlossenen Vokalen (*i:,e:*) dazu führt, dass sie als *offen* (*a:,E:*) klassifiziert werden. Für die Sprachverständlichkeit des potentiellen Lernenden bedeutet diese systematische Schwierigkeit kein Problem, da Sprecher in der Lage sind, die übertriebene Kieferöffnung bei geschlossenen Vokalen mit einer veränderten Zungenstellung zu kompensieren [7].

4 Zusammenfassung & Schlussfolgerung

Es wurde ein Schema für die Bewertung von Sprechbewegungen entwickelt und evaluiert, welches über die Bewertung hinaus bei vorhandenen Defiziten auch zur Erzeugung aussagekräftiger Rückmeldungen an den Lernenden genutzt werden kann.
Basierend auf wiederholten Sprechbewegungen eines Sprechers sowie auf den Sprechbewegungen von insgesamt 14 Sprechern wurde eine Menge von Merkmalen herausgearbeitet, die geeignet ist, sechs Vokalgruppen gegeneinander abzugrenzen. Die Klassifikationsraten für die einzelnen Merkmale sind unterschiedlich hoch; liegen jedoch alle oberhalb von 80%. Die Trennung der Gruppen bezüglich der *Gespanntheit* ist in beiden Experimenten ausgezeichnet. Die Klassifikationsraten bei *Kieferöffnung* und *Lippenrundung* sind dagegen stärker von sprecherspezifischen Faktoren abhängig. Basierend auf der diagonalen Struktur der Verwechslungsmatrix kann dennoch festgestellt werden, dass ein Konzept von korrekter Artikulation in der hier vorgestellten Struktur sinnvoll ist.
Weil unterschiedliche Sprechbewegungen zu gleichermaßen verständlicher Sprache führen, sollte weiterhin mittels moderner Algorithmen der Spracherkennung auf der Ebene des Audiosignals die Entscheidung bezüglich des Erfolgs getroffen werden. Wenn die Erkennung der intendierten Äußerung nicht erfolgreich ist, könnte ein Modul auf der Basis artikulatorischer Merkmale nach dem hier beschriebenen Schema Hinweise zur Korrektur der Sprechbewegung erzeugen.
Das hier verwendete 3D *motion capturing* kann, nicht zuletzt wegen des enormen Aufwandes zur Erfassung und Auswertung der Daten, nicht ohne Weiteres zur Bestimmung artikulatorischer Merkmale im Rahmen selbstständigen Übens verwendet werden. Vielmehr sollte in Zukunft untersucht werden, welche einfacheren Verfahren zur Erfassung der Sprechbewegung verwendet werden können. Die hier evaluierten Merkmale wurden bewusst so definiert, dass eine Bestimmung allein aus der frontalen Ansicht des Sprechers, z.B. mittels digitaler Bilddaten, möglich ist.
Da die Zielgruppe der beschriebenen technischen Systeme größtenteils aus Kindern besteht, wurden auch Bewegungsdaten von Kindern mit dem Ziel analysiert, die Eignung des hier beschriebenen Konzeptes vorerst qualitativ zu beurteilen. Erste Analysen legen nahe, dass nach Anpassung der Schwellwerte innerhalb der ersten Schicht ein Konzept nach dem hier beschriebenen Schema auch als explizite Repräsentation korrekter Sprechbewegungen von Kindern verwendet werden könnte.

5 Danksagung

Diese Arbeit wurde durch Projektmittel der Arbeitsgemeinschaft industrieller Forschungsvereinigungen (AiF) unterstützt, der wir für die finanzielle Förderung (KF 0170802WD7) herzlich danken. Ferner danken wir S. Koppetsch (IB-Hochschule Berlin) und G. Renner (Katholische Hochschule Freiburg) sowie den Kollegen der Phoenix software GmbH Bonn für die Zusammenarbeit im Rahmen des gemeinsamen Projektes.

Literatur

[1] ADAMS, F., H. CREPY, D. JAMESON und J. THATCHER: *IBM products for persons with disabilities*. In: *Communications Technology for the 1990s and Beyond; GLOBECOM 89*, Bd. 2, S. 980–984, 1989.

[2] BÄLTER, O., O. ENGWALL, A.-M. ÖSTER und H. KJELLSTRÖM: *Wizard-of-oz test of ARTUR - a computer-based speech training system with articulation correction*. In: *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility*, S. 36–43, 2005.

[3] CLENDON, S., M. FLYNN und T. COOMBES: *Facilitating speech and language development in children with cochlear implants using computer technology*. Cochlear Implants International, 4(3):119–136, 2003.

[4] H.YEHIA, P. RUBIN und E. VATIKIOTIS-BATESON: *Quantitative association of vocal-tract and facial behavior*. Speech Communication, 26(1-2):23–43, 1998.

[5] IPA: *Handbook of the IPA*. Cambridge University Press, 1999.

[6] KJELLSTRÖM, H. und O. ENGWALL: *Audiovisual-to-articulatory inversion*. Speech Communication, 51:195–209, 2009.

[7] LINDBLOM, B. und J. SUNDBERG: *Acoustical consequences of lip, tongue, jaw, and larynx movement*. Journal of the Acoustical Society of America, 50(4):1166–1179, 1971.

[8] QUINLAN, R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann Pub., 1993.

[9] RENNER, G., S. KOPPETSCH, G.UHLMANN und G.SCHNEIDER: *Sprachspiegel: Nutzung der Spracherkennung zur Sprechtherapie*. Wissenschaftstag des BmWi, 2008.

[10] RIELLA, R., A. LINARTH, L. LIPPERMANN JR. und P. NOHAMA: *Computerized system to aid deaf children in speech learning*. In: *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society*, Bd. 2, S. 1449–1452, 2001.

[11] ROONEY, E., F. CARRARO, W. DEMPSEY, W. DEMPSEY, K. ROBERTSON, R. VAUGHAN, M. JACK und J. MURRAY: *Harp: an autonomous speech rehabilitation system for hearing-impaired people*. In: *Proceedings of the ICSLP*, S. 2019–2022, 1994.

[12] VICSI, K., P. ROACH und A.-M. ÖSTER: *A multimedia, multilingual teaching and training system for children with speech disorders*. International Journal of Speech Technology, 3(3-4):289–300, 2000.

[13] WATSON, C., D. REED, D. KEWLEY-PORT und D. MAKI: *The indiana speech training aid (ISTRA): Comparisons between human and computer-based evaluation of speech quality*. Journal of Speech and Hearing Research, 32:245–251, 1989.