Modell einer Frauenstimme für die artikulatorische Sprachsynthese mit VocalTractLab

Susanne Drechsel[1], *Yingming Gao*[2], *Jens Frahm*[3], *Peter Birkholz*[2]

[1]*Abteilung Sprechwissenschaft und Phonetik, Martin-Luther-Universität Halle*
[2]*Institut für Akustik und Sprachkommunikation, Technische Universität Dresden*
[3]*Biomedizinische NMR, Max-Planck-Institut für biophysikalische Chemie Göttingen*
susannedrechsel@gmx.de

Kurzfassung: Für das artikulatorische Sprachsynthesesystem *VocalTractLab*, das in der veröffentlichten Version auf dem geometrischen Modell eines männlichen Vokaltrakts basiert, wird das Modell für eine Frauenstimme vorgestellt. Anhand von MRT-Aufnahmen, Kieferabdrücken und Sprachaufnahmen einer ausgebildeten Sprecherin wurden die anatomischen Parameter für den weiblichen Vokaltrakt bestimmt und die Zielformen der Einzellaute sowie der glottalen Gesten angepasst. Die Sprachsynthese direkt aus Text oder einer phonetischen Transkription erfolgt mit *VocalTractLab* derzeit noch nicht automatisch. Die Schritte zur Erstellung von gestischen Partituren werden beschrieben und die Ergebnisse einer ersten Hörerbefragung zur Qualität der synthetischen Frauenstimme präsentiert.

1 Einleitung

Synthetische Sprachausgabe begegnet uns in vielen alltäglichen Bereichen. Vorrangig wird dafür im Moment Verkettungssynthese mit verschieden großen natürlichsprachlichen Bausteinen genutzt, von Diphonen über Silben bis hin zu ganzen Wörtern und Phrasen. Je nach Entwicklungsgrad des Inventars können diese Systeme eine sehr hohe bis natürliche Sprachqualität erreichen. Problematisch bleiben jedoch, neben der Prosodiemodellierung, dem System unbekannte Lautverbindungen und hörbare Verkettungsstellen, eine geringe Flexibilität für neue Laute sowie eine geringe Modulationsfähigkeit der Stimmgebung. Die artikulatorische Sprachsynthese verfolgt hingegen die vollständige Simulation des Sprachproduktionsprozesses von der Anregung des Stimmtones durch den glottalen Luftstrom über die akustische Schallausbreitung im modulierten Vokaltrakt bis zur Schallabstrahlung an Mund- und Nasenöffnung. Ein aktuelles, hochqualitatives und frei verfügbares artikulatorisches Synthesesystem ist *VocalTractLab* [1], das in der veröffentlichten Version auf dem Vokaltraktmodell für eine Männerstimme basiert [2].

In einer Kooperation zwischen der Abteilung Sprechwissenschaft und Phonetik der MLU Halle-Wittenberg und dem Institut für Akustik und Sprachkommunikation der TU Dresden wird seit 2002 an der synthetischen Sprachausgabe für das Deutsche Aussprachewörterbuch (DAWB) [3] gearbeitet. Seit der Besetzung der Professur für Kognitive Systeme durch Peter Birkholz im Jahr 2014 kann dafür *VocalTractLab* genutzt werden. Die Ausgabe der rund 150.000 Wörterbucheinträge soll mit einer Frauenstimme umgesetzt werden und sich an der normphonetischen Transkription nach hallescher Tradition orientieren. Die Anforderungen an die Synthese sind hoch: Aussprache und Akzentuierung sollen mustergültig sein und der Stimmklang eine hohe Natürlichkeit aufweisen. Im vorliegenden Beitrag werden die Modellierung des weiblichen dreidimensionalen Vokaltraktmodells beschrieben und das aktuelle Vorgehen zur Erzeugung synthetischer Äußerungen mit *VocalTractLab* vorgestellt.

2 Vokaltraktmodellierung

Zentrales Element der artikulatorischen Synthese ist ein dreidimensionales Vokaltraktmodell. Hier werden die Größenverhältnisse des Ansatzrohres und damit die akustischen Eigenschaften festgelegt. Die Artikulation der einzelnen Laute erfolgt über Parameter, mit denen die Formung und Positionierung der Artikulationsorgane wie Lippenöffnung und -rundung, Höhe und Vor-/Rückverlagerung des Zungenrückens und der Zungenspitze, Gaumensegelöffnung und mehr gesteuert werden. Die Modellierung des weiblichen Vokaltrakts orientiert sich im Wesentlichen an den Methoden von Birkholz 2005 [4] und 2013 [2] sowie Birkholz und Kröger 2006 [5]. Für die Ausmessung der Kieferformen mit Gaumen und Zähnen wurden Gipsabdrücke einer professionellen Sprecherin angefertigt und die übrigen Parameter des Vokaltrakts mit MRT-Scans bestimmt.

2.1 Gipsabdrücke von Ober- und Unterkiefer

Die zahnärztlich angefertigten Gipsabdrücke von Ober- und Unterkiefer der Modellsprecherin wurden mit dem *NextEngine Desktop 3D Scanner 202Oi* maßstabgerecht und hochauflösend abgescannt (Macro, 360°, light, HD). Der Scan erfolgte in zwei Positionen, auf der Unterseite liegend und auf der flachen Rückseite stehend, die Einzelbilder wurden mit *NextEngine ScanStudio* aligniert, zusammengeführt, bearbeitet und im stl-Dateiformat exportiert.

Das 3D-Modell wurde mit der Software *RepSnapper* [6] in Schichten mit einem Abstand von 0,25 mm zerlegt, die als Vektorgrafiken exportiert wurden. Die Konvertierung der Vektorgrafiken in Rastergrafiken erfolgte mit *Adobe Illustrator*. Anhand der Bilddateien wurden in *Image3D* [1] die Konturen von Gaumen und Zähnen nachgezeichnet und ausgemessen (Abbildung 1). Die Maße der Grund- und Kauflächen der Zähne ($w_{t,i}$ und $w_{b,i}$, $i = 0…8$) sowie die Gaumenhöhe entlang mehrerer Punkte der mediosagittalen Kontur wurden in die Sprecherdatei von *VocalTractLab* übertragen, wo die Maße des Drahtgittermodells festgelegt sind (siehe Abbildung 2).

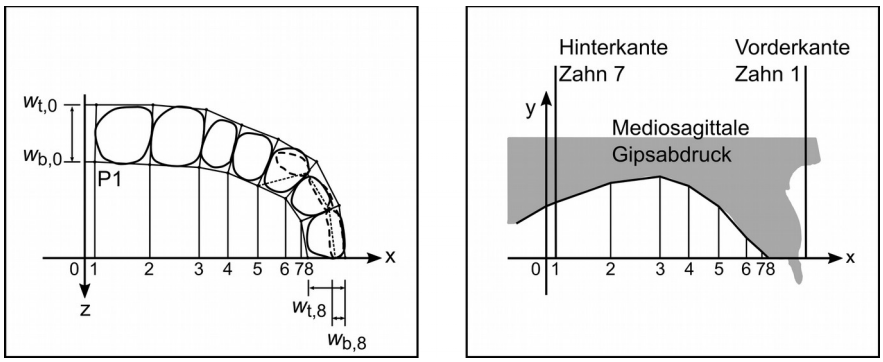

Abbildung 1 - Links: Draufsicht des Oberkiefers mit Zahnkonturen, $w_{b,i}$ und $w_{t,i}$ (width bottom/top) bezeichnen die Flächenbreite der Zähne im Koordinatensystem an der Kaufläche und am Zahnfleisch Rechts: Mediosagittalschnitt des Oberkiefers mit Gaumenhöhe an den Punkten 0 bis 8

2.2 Vokaltrakt-Scans mit Magnetresonanztomographie (MRT)

Für die Modellierung der Größenverhältnisse und Extrempositionen im Vokaltrakt, bspw. der vertikalen Kehlkopfposition oder der Länge des weichen Gaumens, wurden Scans mittels 3D-MRT und Echtzeit-MRT angefertigt [7], [8]. Berücksichtigt wurden alle Laute des deutschen Phonemsystems und die englischen th-Laute.

Mit 3D-MRT wurden alle Laute aufgenommen, die gehalten werden können: die Vokale [aː eː iː oː uː ɛː øː yː ɪ ɔ ʊ ʏ œ ə ɐ], Frikative [f s ʃ ç x θ v z ʒ j ʁ ð], Nasale [m n ŋ] und der Lateral [l]. Jeder Laut wurde 14 Sekunden lang ausgehalten, die Aufnahmen enthalten in der Sagittalebene 36 Scans mit einer Schichtdicke von 3 mm (Messfeld: 224 x 224 mm²).

Anhand der MRT-Bilder wurden für jeden Laut die Konturen der Mediosagittalen nachgezeichnet, wobei auch die Zungenform an den Seitenrändern berücksichtigt wurde (1,2 cm rechts von der Mediosagittalen).

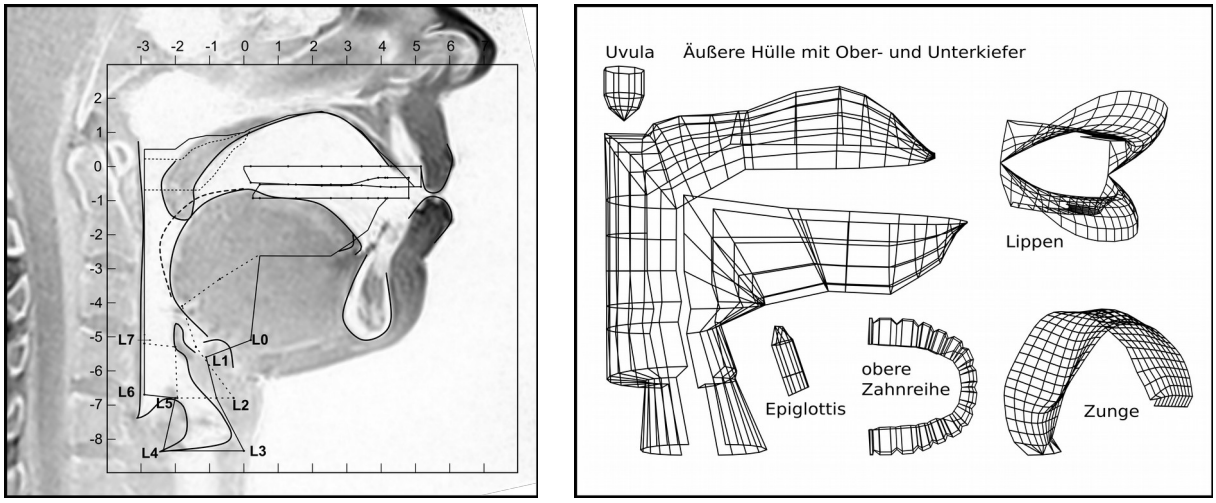

Abbildung 2 - Links: MRT-Bild des Lautes [m] im Sagittalschnitt mit nachgezeichneten Konturen und dem Schema des Vokaltraktes im Koordinatensystem. Rechts: Gitternetz des Vokaltraktes für die Frauenstimme

Um die kontextabhängigen Zielformen für die Konsonanten zu modellieren, wurde ein Korpus an dynamischen MRT-Daten mit einer Schichtdicke von 6 mm und einer Messzeit von 33,3 ms, also 30 Bildern pro Sekunde aufgenommen (Bildauflösung: 1,4 x 1,4 mm², Messfeld: 192 x 192 mm²). Es besteht aus CV-Folgen für alle standarddeutschen Konsonanten inklusive der englischen th-Laute kombiniert mit den gespannten Vokalen [a: i: u:], z. B. *ba-ba-ba, bi-bi-bi, bu-bu-bu*. Mit den Eckvokalen werden die Extrempositionen der Artikulatoren abgebildet: mit [a:] die maximale Kieferöffnung/Zungensenkung, mit [i:] die maximale Zungenhebung im vorderen Drittel des Mundraums und mit [u:] die maximale Zungenhebung im hinteren Drittel sowie die Lippenrundung. Nach der visuellen Parameterbestimmung anhand der nachgezeichneten Konturen für die Vokaltrakteinstellungen je Laut wurden akustische Optimierungen mit den für die Sprecherin gemessenen Formanten F_1-F_3 in *VocalTractLab* durchgeführt, vgl. [2]. Gegebenenfalls wurden die Vokaltraktformen an eine gemäßigte Artikulation angepasst, z. B. die Kieferöffnung für /a/, die von der Sprecherin im MRT zu extrem umgesetzt wurde.

2.3 Anpassung des geometrischen Glottismodells

In der hier verwendeten Version von *VocalTractLab* wird das Quellsignal durch ein geometrisches Glottismodell (noch nicht veröffentlicht) erzeugt, ähnlich dem Modell von Titze 1984 [9]. Physiologisch betrachtet ist der männliche Kehlkopf etwa 20 % größer als der weibliche. Daraus resultieren Unterschiede in der Stimmlippenlänge, die sich auf die Indifferenzlage der Sprechstimme auswirken. Bei Männern liegt die Indifferenzlage (statistisch definierter Normbereich) zwischen 87-131 Hz, bei Frauen zwischen 175-262 Hz [10]. Die Angaben der mittleren Stimmlippenlänge sind breit gestreut, da es bislang keine präzise Messmethode gibt. Titze 1989 [11] gibt für Männer 16 mm und für Frauen 10 mm an; Nawka/Wirth 2008 [10] geben für eine Bassstimme 24-25 mm an und für eine Sopranstimme 14-17 mm. Gleich bleibt das Verhältnis: Männliche Stimmlippen sind 1,6 Mal länger als weibliche. Für den hinteren knorpligen Teil gilt das nicht; hier gibt es fast keine Längendifferenz [11].

Tabelle 1 **- Statische Parameter für das geometrische Glottismodell in *VocalTractLab***

statische Parameter	**männlich**	**weiblich**
Rest thickness	4.5 mm	4 mm
Rest length	16 mm	12 mm
Rest f_0	120 Hz	200 Hz
Chink length	4.0 mm	4.0 mm

Für die Synthesestimme wurden die Maße des geometrischen Glottismodells wie in Tabelle 1 angepasst. Auch die Formen für die glottalen Gesten wurden verändert. Durch den bei Frauenstimmen häufig zu beobachtenden (physiologischen) unvollständigen Stimmlippenschluss im hinteren Drittel entsteht ein leicht behauchter oder weicherer Stimmklang. Um diesen Effekt in der Frauenstimme zu modellieren, wurden die glottalen Zielformen angepasst. In Tabelle 2 sind beispielhaft die Parameterwerte für die modale Stimmgebung im Vergleich von der synthetischen Männer- zur Frauenstimme angegeben.

Tabelle 2 **- Parameter in *VocalTractLab* für die modale Glottisform im Vergleich männl./weiblich**

Gestenform	Lower displ.	Upper displ.	Chink area	Phase lag	Relat. amplitude	Double pulsing	Pulse skewness	Aspir. strength
modal_male	0.3 mm	0.3 mm	2 mm²	50 deg	1	0	0	-40 dB
modal_female	0.6 mm	0.6 mm	4 mm²	50 deg	0.8	0	0.3	-27 dB

3 Erstellung gestischer Partituren

Die Ansteuerung der Artikulatoren erfolgt in *VocalTractLab* über gestische Partituren. Hier wird die Abfolge und Geschwindigkeit der Artikulationsbewegungen und damit die Dauer der erzeugten Sprachlaute geregelt. Eine Partitur enthält acht Spuren, in den oberen fünf werden die Bewegungen der Artikulationsorgane gesteuert, die unteren sind für Modi der Stimmgebung (Glottalplosiv, Behauchung, Flüstern etc.), Grundfrequenzbewegung und Luftdruck bestimmt (s. Abbildung 3). Die einzelnen Gestenintervalle beinhalten die Parameter Zielposition, Geschwindigkeit der Bewegung (Zeitkonstante) und Dauer.

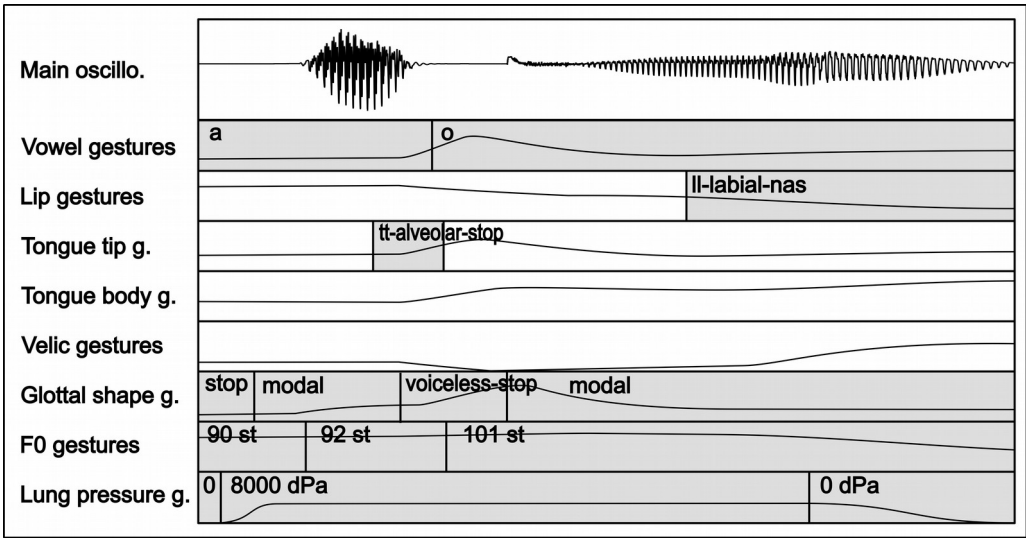

Abbildung 3 - Partitur für das Wort „Atom“

Um Partituren automatisch aus der Transkription zu erstellen, gab es bisher verschiedene Ansätze. Birkholz et al. [12] entwickelten Regeln für die Übersetzung von phonetischen Symbolen in Gestenintervalle und nutzten dabei die Lautdauervorhersage des *Bonn Open Synthesis System* (BOSS). Problematisch ist, dass die Gestendauern nicht den Lautdauern entsprechen. Deshalb nutzten sie in einem zweiten Ansatz Daten aus elektromagnetischer Artikulographie (EMA), um natürlichsprachliche Äußerungen zu resynthetisieren. Beide Ansätze lieferten für Wörter mit einfacher CV-Struktur verständliche Ergebnisse.

Weitz et al. [13] integrierten *VocalTractLab* in das Text-to-Speech-System *MARY* und entwickelten einen regel- und einen datenbasierten Ansatz, um die phonetische Repräsentation in Gesten zu übersetzen. Die Resultate aus der regelbasierten Synthese waren verständlich, klangen aber aufgrund der unzulänglichen Lautdauern unnatürlich. Mit den statistischen Modellen konnten meist nur unverständliche Resultate erzielt werden.

Bei der Synthese für ein Aussprachewörterbuch müssen komplexe Silbenstrukturen berücksichtigt werden, da das Deutsche im Silbenonset bis zu drei und in der Silbencoda bis zu fünf Konsonanten aufweisen kann – je nachdem, ob flektierte Formen einbezogen werden. Es gibt noch keine Möglichkeit, Partituren vollständig automatisch zu generieren und dabei gute Syntheseergebnisse zu erzielen. Deshalb wurden für die Bewertung durch Testhörer Stimuli in drei Schritten erstellt: zunächst die Übersetzung der Transkription in initiale Gestenpartituren, dann die Erzeugung der Grundfrequenzkontur anhand natürlichsprachlichen Materials und abschließend die manuelle Korrektur.

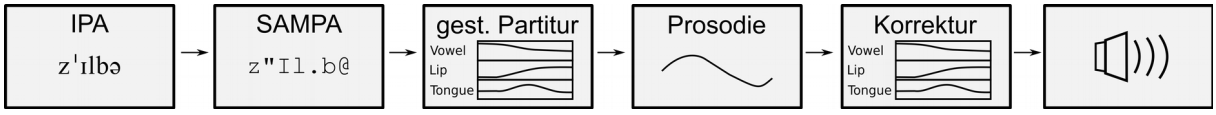

Abbildung 4 - Einzelschritte von der Transkription zur synthetischen Äußerung

3.1 Initiale Partituren aus SAMPA-Transkriptionen

Für die Erstellung von initialen Partituren wurde ein Skript entwickelt, das SAMPA-Symbole in Gestenintervalle übersetzt. Dafür wurden das Zeitstrukturmodell von Xu adaptiert [14] und die inhärenten Lautdauern für das Deutsche von Kohler [15] übernommen. Die Organisation der Gesten erfolgt silbenweise, jede Silbe muss in der Transkription mit einem Konsonanten oder einem Glottalplosiv beginnen. Alle Gesten am Beginn jeder Silbe sind zeitlich aligniert; Vokal-, Konsonant- und Glottisgeste beginnen gleichzeitig. Konsonantengesten in der Silbencoda beginnen am Ende der vokalischen Geste. Die Dauer von glottalen Gesten für z. B. stimmlose Konsonanten entspricht der Gestendauer des Artikulators, die Zeitkonstante wurde für alle Modi auf 15 ms festgesetzt. Die Zeitkonstanten für Artikulatoren wurden laut Birkholz et al. [16] mit 10 ms für Lippengesten, 15 ms für Zungenspitzengesten (davon abweichend 5 ms für den Lateral) und 20 ms für Zungenkörpergesten festgelegt. Die Zeitkonstante für den Vokal einer Silbe entspricht der des Onsetkonsonanten. Da *VocalTractLab* auf dem *Target Approximation Model* basiert und die artikulatorischen Ziele graduell erreicht werden, verzögern sich die tatsächlichen Artikulationsbewegungen im Verhältnis zu ihren Gesten (siehe Abbildung 3). Deswegen wird am Beginn der Zeile für den Lungendruck ein leeres Intervall eingefügt, dessen Dauer vom Artikulationsmodus des Onsetkonsonanten abhängt: 100 % der Gestendauer für Plosive, 50 % für Frikative und 30 % für Sonoranten. Zeitlich aligniert mit der letzten Lautgeste wird ein Extraintervall von 100 ms Dauer für den Lungendruck eingefügt, um den Luftstrom bis zum Äußerungsende zu gewährleisten. In den initialen Partituren ist noch keine f_0-Kontur vorhanden.

3.2 Prosodiemodellierung

Die Gesten für den Grundfrequenzverlauf wurden mit der Software *TargetOptimizer* [1], [17] bestimmt. Dazu wurde das Korpus der Hörbeispiele, die dem DAWB beiliegen [18], in Praat [19] silbenweise segmentiert und die TextGrid- sowie die PitchTier-Dateien (spreadsheet file) gespeichert. Auf Grundlage dieser Dateien können im *TargetOptimizer* die Grundfrequenzparameter Dauer (s), Zielfrequenz (st), Anstieg (st/s) und Zeitkonstante (ms) je Silbe eines Wortes automatisch bestimmt und als Textdatei oder gestische Partitur exportiert werden.

3.3 Manuelle Korrektur

Die Optimierung der initialen Partitur erfolgte durch Resynthese der originalen Äußerungen. Dazu wurden jeweils die initiale Partitur, Grundfrequenz-Partitur und das eingesprochene Zielwort in *VocalTractLab* geladen und manuell aufeinander abgestimmt. Die glottalen Gesten sind abhängig von den Ziellauten anzupassen. Zum Beispiel wird die Aspiration der stimmlosen Plosive erzeugt, indem die glottale Öffnungsgeste etwa ein Drittel länger ist als die Geste des verschlussbildenden Artikulators. Die Zeitkonstante für das letzte Lungendruck-Intervall wurde individuell erhöht, sodass der letzte Laut ausschwingen kann und die finale Dehnung erreicht wird.

4 Hörerbewertung

Für die qualitative Bewertung der Frauenstimme wurden nach dem oben beschriebenen Verfahren Stimuli generiert. Dafür wurden aus dem Korpus der DAWB-Hörbeispiele 20 Wörter ausgewählt, mit denen alle Laute des deutschen Phonemsystems inklusive der Diphthonge [aɛ̯], [aɔ̯] und [ɔœ̯] abgedeckt sind. Die Wörter haben unterschiedlich komplexe Silbenstrukturen und sind zwei-, drei-, oder fünfsilbig (siehe Tabelle 3). Die Stimuli wurden über Praat und mit einem Bluetooth-Lautsprecher der Marke *Ultimate Ears Boom 2* abgespielt.

Die synthetische Frauenstimme wurde von 23 Studierenden des Masterstudiengangs der Sprechwissenschaft bewertet (Altersdurchschnitt: 26). Zuerst wurde die Verständlichkeit geprüft, indem jedes Wort zweimal vorgespielt wurde und das Verstandene notiert werden sollte. Die Wörter wurden im Mittel zu 84 % richtig erkannt. Elf Wörter wurden zu 100 % richtig erkannt. Das niederfrequente Wort „Gespött“ wurde nur sechs mal richtig erkannt, im Wort „wiederaufbauen“ wurde häufig der Laut [b̥] als [t] verstanden.

Tabelle 3 **- Stimuli mit Erkennungsrate**

heute, Idee, Champagner, empfinden, Garage, Museum, Österreich, unmöglich, Ursache, Universität, Wiedergutmachung					**100 % richtig erkannt**	
	richtig erkannt		**Einzellaut falsch**		**nicht erkannt**	
süddeutsch	22	96 %	0	0 %	1	4 %
Informatiker	21	91 %	2	9 %	0	0 %
Unglück	21	91 %	1	4 %	1	4 %
Eleganz	20	87 %	1	4 %	2	9 %
Gymnastik	17	74 %	5	22 %	1	4 %
missglücken	17	74 %	0	0 %	6	26 %
Atom	7	30 %	0	0 %	16	70 %
Gespött	6	26 %	0	0 %	17	74 %
wiederaufbauen	4	17 %	14	61 %	5	22 %
MW Gesamt:	84%		5 %		11%	

Die Synthesestimme sollte außerdem in Hinblick auf die Verwendung für ein Aussprachewörterbuch bewertet werden. Zwei Drittel der Befragten lehnen die Synthesestimme für diesen Zweck ab, weil noch zu viele Unsicherheiten auf lautlicher Ebene vorhanden sind und die Stimme nicht angenehm genug klingt. Als Orientierung für eine normphonetische Aussprache bedarf die Synthese insgesamt noch der Verbesserung, auch wenn einzelne Wörter schon als gut bis sehr gut bewertet wurden. Die Prosodie wird einerseits als monoton, andererseits als zu extrem akzentuiert beschrieben. Auf einer fünfstufigen Skala (1 = sehr gut, 2 = gut, 3 = akzeptabel, 4 = eher schlecht und 5 = schlecht) wurde die Frauenstimme im Mittel mit 3,4 (Medianwert 3) bewertet.

5 Zusammenfassung und Ausblick

In diesem Beitrag wurde eine deutsche Frauenstimme mit artikulatorischer Sprachsynthese vorgestellt. Das Vorgehen zur Modellierung des Vokaltraktes und ein Ansatz zur Erstellung von gestischen Partituren wurden beschrieben. Die Synthese-Ergebnisse sind weitgehend verständlich und Studierende der Sprechwissenschaft bewerten die synthetische Frauenstimme im Mittel als akzeptabel. Die Güte auf segmentaler Ebene hängt stark von dem Zusammenspiel der einzelnen Gesten ab und kann bei komplexeren Silbenstrukturen noch nicht durch ausschließlich automatische TTS gewährleistet werden. Für die Prosodiemodellierung wird auf natürlichsprachliches Material zurückgegriffen.

Perspektivisch soll das gesamte Wörterbuchkorpus synthetisiert werden. Dazu ist es erforderlich, dass auch komplexe sowie eingedeutschte fremde Wörter und Wortgruppen mit guter Qualität automatisch generiert werden können. Das Skript zur Erstellung initialer Gesten soll sukzessive verfeinert werden, indem weitere Regeln implementiert werden. Zum Beispiel kann das Verhältnis von artikulatorischen und glottalen Gesten für intervokalische stimmlose Plosive oder die erforderlichen Gestenfolgen für Konsonanten-Häufungen vorbestimmt werden. Die Prosodie könnte anhand von Akzentstufen, die aus der Transkription ablesbar sind, modelliert werden. Im Akzentstufenmodell erhalten Silben mit Hauptakzent die Stufe 4, Silben mit Nebenakzent die Stufe 3, unbetonte Silben Stufe 2 und reduzierte Silben Stufe 1 [20]. Versuche mit statistischer Vorhersage, die das Akzentstufenmodell einbeziehen, haben jedoch noch keine zufriedenstellenden Ergebnisse gebracht, was in der geringen Datenbasis begründet ist [21]. Auf Grundlage der Befragungsergebnisse wird die Artikulationsbasis und die wahrgenommene Stimmqualität weiter verbessert.

Literatur

[1] www.vocaltractlab.de

[2] BIRKHOLZ, P.: *Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis.* In: PLoS ONE, 8(4): e60603, 2013. (doi:10.1371/journal.pone.0060603)

[3] KRECH, E.-M., E. STOCK, U. HIRSCHFELD, L. C. ANDERS: *Deutsches Aussprachewörterbuch*. De Gruyter Berlin, New York, 2009.

[4] BIRKHOLZ, P.: *3D-Artikulatorische Sprachsynthese*. Logos Verlag, Berlin, 2005.

[5] BIRKHOLZ, P., B. KRÖGER: *Vocal Tract Model Adaptation Using Magnetic Resonance Imaging*. In: *Proc. of the 7th International Seminar on Speech Production 2006*, S. 493–500, Ubatuba, Brazil, 2006.

[6] RepSnapper 2.2.0, https://reprap.org/wiki/RepSnapper_Manual:Introduction

[7] Uecker, M., S. Zhang, D. Voit, A. Karaus, K. D. Merboldt, J. Frahm: *Real-time MRI at a resolution of 20 ms.* In: *NMR in Biomedicine* 23, S. 986–994, 2010. (doi: 10.1002/nbm.158)

[8] Niebergall, A., S. Zhang, E. Kunay, G. Keydana, M. Job, M. Uecker, J. Frahm: *Real-time MRI of speaking at a resolution of 33 ms: Undersampled radial FLASH with nonlinear inverse reconstruction.* In: *Magnetic Resonance in Medicine* 69, S. 477–485, 2013. (doi: 10.1002/mrm.24276)

[9] Titze, I.: *Parameterization of the glottal area, glottal flow, and vocal fold contact area*. In: *The Journal of the Acoustical Society of America*, 75, S. 570–580, 1984.

[10] Nawka, T., G. Wirth: *Stimmstörungen. Für Ärzte, Logopäden, Sprachheilpädagogen und Sprechwissenschaftler.* Dt. Ärzte-Verlag, Köln, 2008.

[11] Titze, I.: *Physiological and acoustic differences between male and female voices.* In: *Journal of the Acoustical Society of America*, 85, S. 1699–1707, 1989.

[12] Birkholz, P., I. Steiner, S. Breuer: *Control concepts for articulatory speech synthesis*. In: *Proceedings of the 6th ISCA Speech Synthesis Research Workshop 2007*, S. 5–10, Bonn 2007.

[13] Weitz, B., I. Steiner, P. Birkholz: *Gesture-Based Articulatory Text-to-Speech Synthesis*. In: J. Trouvain, I. Steiner, B. Möbius (Hrsg.): *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2017*, Band 86, S. 324–331. TUDPress, Dresden, 2017.

[14] Xu, Y., F. Liu: *Tonal alignment, syllable structure and coarticulation: Toward an integrated model.* In: *Italian Journal of Linguistics*, 18(1), S. 125–159, 2006.

[15] Kohler, K. J.: *Zeitstrukturierung in der Sprachsynthese*. In: A. LACROIX (Hrsg.): *ITG-Tagung Digitale Sprachverarbeitung*, S. 165–170. vde-Verlag, Berlin, Bad Nauheim, 1988.

[16] Birkholz P, L. Martin, Y. Xu, S. Scherbaum, C. Neuschaefer-Rube: *Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis*. In: *Computer Speech & Language*, 41, S. 116–127, 2017. (doi: 10.1016/j.csl.2016.06.004)

[17] Birkholz, P., P. Schmager, Y. Xu: *Estimation of Pitch Targets from Speech Signals by Joint Regularized Optimization*. In: *26th European Signal Processing Conference (EUSIPCO)*, S. 2089–2093, 2018.

[18] Download Hörbeispiele DAWB: www.degruyter.com/view/product/19839

[19] www.fon.hum.uva.nl/praat/

[20] Drechsel, S.: *Aufbereitung des Halle-Korpus für die maschinelle Verarbeitung*. In: A. Ebel (Hrsg.): *Beiträge zum 5. Doktorandentag der Halleschen Sprechwissenschaft.* Online-Publikation in Vorbereitung.

[21] Schmager, P.: *Vergleich maschineller Lernverfahren für die Grundfrequenzvorhersage in der Sprachsynthese.* Unveröffentlichte Diplomarbeit, TU Dresden, 2017.