Ein System zur Bildanalyse des menschlichen Vokaltraktes

Johannes Behrends[1], Axel Wismüller[1], Phil Hoole[2]
[1]*Institut für Klinische Radiologie,* [2]*Institut für Phonetik und Sprachliche Kommunikation, LMU München,*
johnny_b@gmx.li, axel@wismueller.de, hoole@phonetik.uni-muenchen.de

Einleitung

Die Magnetresonanztomographie (MRT) als modernes Schnittbildverfahren in der medizinischen Diagnostik ermöglicht die *in vivo* Analyse der dreidimensionalen Struktur des menschlichen Vokaltraktes. Innovative Methoden zur computerunterstützten Bildanalyse und automatischen Mustererkennung ermöglichen hieraus die Gewinnung akustisch-artikulatorischer Modelle als einen wesentlichen Beitrag zur Grundlagenforschung in der Phonetik und zur Konstruktion technischer Systeme für die Sprachsynthese [1, 2, 3, 4]. Hierzu werden MRT-Untersuchungen während des Phonationsaktes in unterschiedlicher Schnittführung durchgeführt. Die anatomisch korrekte Registrierung dieser Daten ermöglicht die hochpräzise 3D-Rekonstruktion des Vokaltrakts in isotroper Auflösung. Nach halbautomatischer Segmentierung zur Extraktion des Ansatzrohres wird durch eine nichtlineare Hauptachsentransformation eine gekrümmte Mittellinie errechnet, wodurch der Vokaltrakt durch eine Kaskade von Zylindern charakteristischer Querschnittsfächen als *Areafunktion* approximiert werden kann. Die in vivo MRT-Analyse des menschlichen Phonationsaktes liefert somit wertvolle Erkenntnisse an der Nahtstelle zwischen akustischer Phonetik, medizinischer Diagnostik und digitaler Bildverarbeitung. Für das Erstellen hochpräziser akustisch-artikulatorischer Modelle des menschlichen Vokaltraktes sind detaillierte Kenntnisse über dessen dreidimensionale Beschaffenheit erforderlich. Ein solches Modell wird gewonnen, indem man die Querschnittsflächen des Vokaltraktes entlang einer Mittellinie bestimmt und ihn als eine Kaskade von Zylindern dieser Querschnittsflächen (*Areafunktion*) annähert. Ziel dieser Arbeit ist es, ein Verfahren zu entwickeln, mit dem der Vokaltrakt auf diese Weise mit möglichst geringem menschlichen Aufwand modelliert werden kann. Es wird dabei sowohl auf Fragestellungen der Bildverarbeitung als auch der akustischen Phonetik Bezug genommen.

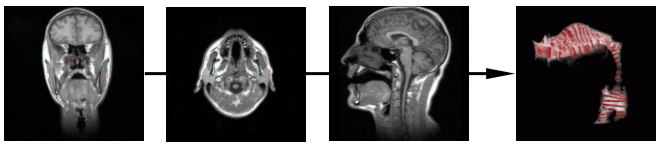

Abbildung 1: Zielsetzung: Aus 3D MRT Datenmaterial soll der Vokaltrakt extrahiert und seine Querschnittsflächenfunktion ermittelt werden.

Daten

Von neun trainierten Sprechern wurden 3D MRT-Datensätze (T1-gewichtete Sequenz, Matrixgröße 256×256, Voxelgröße $1.17\times1.17\times5\text{mm}^3$) in sagittaler (15 Schnitte), koronarer und transversaler (jeweils 23 Schnitte) Schichtführung während der Produktion von Lauten gewonnen. Die Sprecher gaben während der Aufname Laute aus dem deutschen vokalinventar (/a/, /e/, /i/, /o/, /u/, /y/, /ø) sowie die alveolaren Konsonanten /s/, /S/ (sprich: „sch“), /n/ und /l/ wieder. Das Sprachsignal wurde zwei Sekunden vor und während der gesamten Messung aufgezeichnet, um die Korrektheit der Aussprache zu kontrollieren. Weiterhin wurden von jedem Sprecher Gebißabdrücke angefertigt, von denen hochauflösende computertomographische (CT) 3D Datensätze gewonnen wurden (Matrixgröße $512\times512\times200$, Voxelgröße $0.156\times0.156\times0.3\text{mm}^3$), ohne den Sprecher selbst der Röntgenstrahlung auszusetzen.

Methoden

Bildüberlagerung

Durch die hohe Schichtdicke und der daraus resultierenden Anisotropie der Voxel wird die Genauigkeit des Modells im auf die räumliche Auflösung in z-Richtung beeinträchtigt. Eine einfache Neuabtastung des Datensatzes auf isotrope Voxel höchstmöglicher Auflösung würde auch nach Glättung in z-Richtung auf Stufenartefakte bei der 3D-Rekonstruktion führen.

Durch die Bildakquisition in drei orthogonalen Schichtführungen (s. o.) kann der Informationsverlust eines Datensatzes in z-Richtung durch den Informationsgehalt eines anderen Datensatzes in x- oder y-Richtung ausgeglichen werden. Dies geschieht durch eine anatomisch korrekte Überlagerung und Mittelwertbildung aller drei Datensätze durch den *Automated Image Registration (AIR)*-Algorithmus [5] (vgl. Abbildung 2).

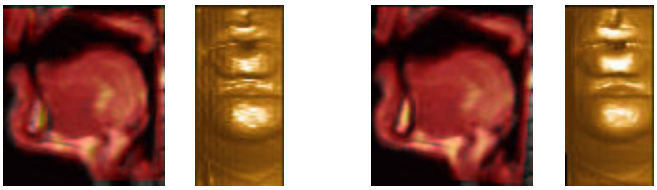

Abbildung 2: Überlagerung zweier orthogonaler Datensätze vor (links) und nach (rechts) AIR. Der Vergleich der Oberflächendarstellungen verdeutlicht die Qualitätsverbesserung.

Segmentierung

Grundlage der Segmentierung ist das 3D Region Growing, das für diese Anwendung bereits in [4] verwendet wurde. Hierfür ergeben sich folgende Problemzonenm, durch die das Region Growing in unerwünschte Bereiche hineinlaufen kann (vgl. Abbildung 3): (i) Natürliche Öffnungen des Vokaltraktes, wie die Mundöffnung und die Stimmritze und (ii) solche, die durch unzureichende Abbildungseigenschaften der MRT einen Luft-ähnlichen Grauwert repräsentieren, also die Zähne und der harte Gaumen. (i) kann gelöst werden,

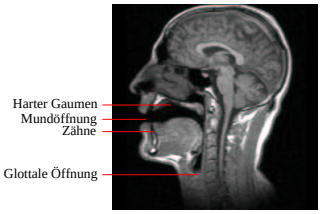

Abbildung 3: Problemzonen für das Region Growing.

indem die CT-Aufnahmen der Gebißabdrücke anatomisch korrekt mit den MRT-Daten überlagert werden (vgl. Abbildung 4a). Das Verschließen der Mundöffnung und der Stimmritze erfolgt durch einen referenzpunkt-basierten Ansatz. Nach einer schichtweisen 2D-Faltung der Mundöffnung mit einem I-förmigen Faltungskern (d. h. die Mundöffnung wird vertikal verschmiert) können Kopfmasken mittels 3D Region Growing geewonnen werden. Die glottale Öffnung wird verschlossen, indem ein Referenzpunkt dicht unterhalb der Glottis gesetzt wird, wodurch alle unterhalb gelegenen Bildbereiche von der Vokaltrakt-Segmentierung ausgeschlossen werden (vgl. Abbildung 4a-c).

Der Vokaltrakt kann nun vom umgebenden Gewebe getrennt werden, indem man einen Punkt in die Vokaltrakt-Region setzt, von dem aus ein weiteres Region Growing startet (vgl. Abbildung 4d).

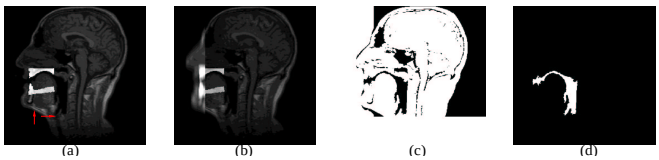

Abbildung 4: Vokaltrakt-Segmentierung für den Vokal /i/.

Mittellinie und Areafunktion

Da konventionelle zweidimensionale Verfahren zur Berechnung der Mittellinie auf Basis der Mediosagittalen Schicht Asymmetrien des Vokaltraktes nicht Rechnung tragen, kann hiermit offenbar eine realistische Evaluation der funktionelle Anatomie während der Phonation nicht durchgeführt werden. Um diese Problematik zu überwinden wurde mittels eines Neuronalen Netzes – einer modifizierten selbstorganisierenden Merkmalskarte (SOM) mit eindimensionaler Topologie [7, 6] – ein Verfahren zur Ermittlung einer gekrümmten 3D-Mittellinie entwickelt (vgl. Abbildung 5).

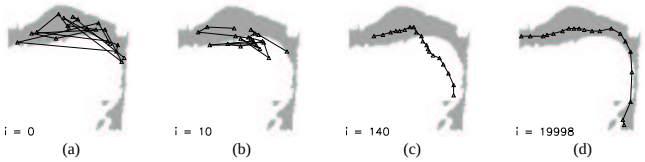

Abbildung 5: Prinzip des modifizierten SOM-Algorithmus: Nach zufälliger Initialisierung verteilt sich die Kette immer mehr innerhalb des Vokaltrakts.

Um ein Überfalten der Topologie zu vermeiden, wurde der SOM-Algorithmus modifiziert, indem die lokale Nachbarschaft σ_r jedes Neurons in der Umgebung ihres kritischen Wertes σ_r^c, bei dem die Überfaltung geschieht, gehalten wird [6].

Definiert man $\alpha = |r' - r''|$ als den Abstand zwischen dem nächsten Neuron r' und dem zweitnächsten Neuron r'' bzgl. dem aktuellen Datenpunkt, beobachtet man eine Topologieverletzung, wenn $\alpha > 1$. In diesem Fall wird σ_r lokal erhöht mittels

$$\sigma_r := \max\left(\sigma_r, \alpha K \exp\left(-\frac{2(r-R)^2}{\alpha^2}\right)\right), \; R = \frac{1}{2}(r' + r''),$$

wobei K empirisch ermittelt wird.

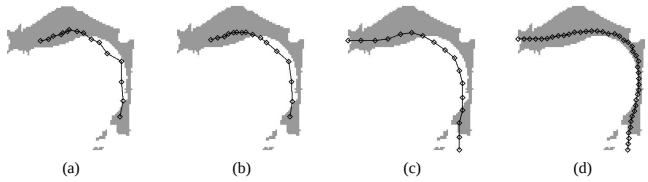

Abbildung 6: Konstruktion der endgültigen Mittellinie.

Für die Konstruktion der endgültigen Mittellinie wird die eindimensionale SOM-Kette C verwendet als Initialisierung für einige Nachverarbeitungsschritte (vgl. Abbildung 6a), die Glättung und Extrapolation beinhalten: (i) Glättung von C mit einem Faltungskern, der mit wachsender Nachbarschaft abnimmt. Das Resultat ist ein geglättetes Codebook $\tilde{C}$ (vgl. Abbildung 6b). (ii) Extrapolation von $\tilde{C}$ in Richtung der Glottis bzw. der Mundöffnung und Neuabtastung auf $\tilde{N}$ äquidistante Punkte $\tilde{P}$ (vgl. Abbildung 6c). (iii) Berechnung von Normalenvektoren $\tilde{\mathbf{n}}_i$ eines jeden Punktes $\tilde{\mathbf{p}}_i$ in $\tilde{P}$ mittels $\tilde{\mathbf{n}}_i = \tilde{\mathbf{p}}_{i-1} - \tilde{\mathbf{p}}_{i+1}$. Diese Normalenvektoren stehen senkrecht auf einer schiefen Schnittebene $\tilde{S}_i$ durch den Vokaltrakt. Für die Randpunkte $\tilde{\mathbf{p}}_1$ $\tilde{\mathbf{p}}_{\tilde{N}}$ wird $\tilde{P}$ extrapoliert. (iv) Berechnung von $\tilde{\mathbf{q}}_i$ als Schwerpunkt der korrespondierenden Querschnittsfläche $\tilde{S}_i$ durch den Vokaltrakt. Daraus resultiert eine Kurve $\tilde{Q}$, die wiederum auf äquidistante Punkte abgetastet wird. (v) Glättung der Kurve $\tilde{Q}$ durch erneute Anwendung von Schritt (i) zur endgültigen Mittellinie Q (vgl. Abbildung 6d).

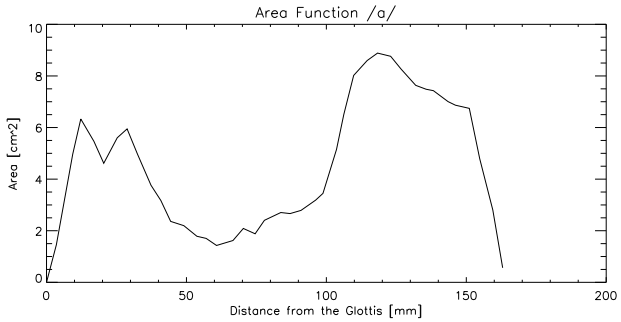

Abbildung 7: Areafunktion für den Vokal /a/; die Querschnittsfläche des Vokaltraktes ist über dem Abstand zur Glottis aufgetragen.

Die Ermittlung des Inhalts der Querschnittsfläche an jedem Stützpunkt führt auf die Areafunktion (vgl. Abbildung 7).

Zusammenfassung und Ausblick

Die MRT in unterschiedlicher Schichtorientierung ermöglicht nach anatomisch korrekter Registrierung eine schnelle und präzise in vivo Evaluation des menschlichen Vokaltraktes während der Phonation. Anhand dieser Information können akustisch-artikulatorische Modelle mittels computergestützter Bildanalysemethoden gewonnen werden. In diesem Zusammenhang trägt die Berechnung einer gekrümmten 3D Mittellinie durch den Vokaltrakt auf der Grundlage selbstorganisierender Merkmalskarten den Asymmetrien des Vokaltraktes Rechnung. Auf diese Weise können die Resultate mit Hinblick auf die Areafunktion gegenüber mediosagittalen 2D Methoden verbessert werden. Diese Evaluation kann interessante Beiträge zur Konstruktion von Sprachsynthese-Systemen liefern und langfristig hilfreich sein in der Analyse von angeborenen oder durch einen chirurgischen Eingriff erworbenen Abnormalitäten, die die funktionelle Anatomie des Oropharynx während der Lautproduktion betreffen.

[1] G. Fant. *Acoustic Theory of Speech Production*. Mouton, den Haag, 1960.

[2] P. Mermelstein. Articulatory Model for the Study of Speech Production. *Journal of the Acoustical Society of America*, 53(4):1070–1082, 1973.

[3] T. Baer, J.C. Gore, and R.C. Gracco. Analysis of Vocal Tract Shape and Dimension using Magnetic Resonance Imaging: Vowels. *Journal of the Acoustical Society of America*, 90(2):799–828, 1991.

[4] I. Titze and B. Story. Vocal Tract Area Functions from Magnetic Resonance Imaging. *Journal of the Acoustical Society of America*, 100(1):537–554, 1996.

[5] R.P. Woods, S.R. Cherry, and J.C. Mazziotta. Rapid Automated Algorithm for Aligning and Reslicing PET Images. *Journal of Computer-Assisted Tomography*, 16:620–633, 1992.

[6] R. Der and M. Hermann. Second-Order Learning in Self-Organizing Maps. In *Kohonen Maps*. E. Oja, S. Kaski, 1999.

[7] T. Kohonen. *Self-Organizing Maps*. Springer, Heidelberg, New York, 2001.