Artikulatorische Sprachsynthese mit dem Programm *TractSyn* - Ein Überblick

Peter Birkholz, Dietmar Jackél

Institut für Informatik, Universität Rostock, Email: piet@informatik.uni-rostock.de

Einleitung

Dieser Artikel stellt den gegenwärtigen Entwicklungsstand des artikulatorischen Sprachsynthetisators *TractSyn* vor. Das System basiert auf einem dreidimensionalen, geometrischen Modell des Vokaltrakts, einem gestisch motivierten Verfahren zur Steuerung der Artikulation und einer zeitdiskreten, aerodynamisch-akustischen Simulation der Luftbewegung im Sprechtrakt. In den folgenden Abschnitten wird jede der Komponenten näher vorgestellt.

Modell des Sprechapparats

Der gesamte Sprechapparat, bestehend aus den subglottalen Luftwegen, der Glottis, dem Vokaltrakt und dem Nasenraum, wird durch die Aneinanderreihung kurzer Rohrabschnitte mit jeweils konstantem Querschnitt modelliert. Es handelt sich also um die klassische, stückweise konstante Approximation der Rohrquerschnittsfunktion. Diese ist in Abb. 1 für den Vokal /o:/ dargestellt. Die

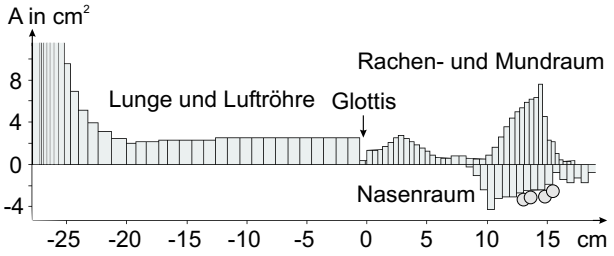

Abbildung 1: Rohrmodell des Sprechapparats.

Querschnittsfunktion des Rachen- und Mundraums wird aus einem 3D-Modell des Vokaltrakts gewonnen, welches eine Weiterentwicklung unseres früheren Modells (siehe [1]) ist. Abb. 2 zeigt die Geometrie des Modells für den Vokal /a:/. Das Modell besitzt 21 Parameter, die direkt die Form und die Lage der Artikulatoren bestimmen: 2 für das Zungenbein, 3 für den Unterkiefer, 2 für die Lippen, 1 für das Gaumensegel und 12 für die Zunge. Vier der Zungenparameter bestimmen die Höhe der Zungenränder gegenüber der Zungenmitte, so dass auch laterale Passagen und mediosagittale Vertiefungen modelliert werden können. Die 3D-Gemoetrie wird in das eindimensionale Rohrmodell überführt, indem die Querschnittsflächen des Vokaltrakts entlang der Schallausbreitungslinie berechnet werden. Im Bereich der Zungenspitze, Zähne, und Lippen erfolgt eine feinere Abtastung der Querschnittsflächen als im Rest des Vokaltrakts. Die feinere Auflösung der Querschnittsfunktion in diesem Bereich hat sich für die Synthese der Frikative als vorteilhaft herausgestellt, da ihr Spektrum empfindlich von der genauen Rohrgeometrie stromabwärts von der Konstriktion abhängt.

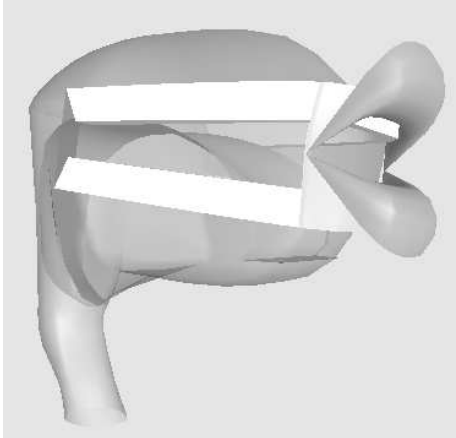

Abbildung 2: Dreidimensionales Modell des Vokaltrakts.

Die Glottis wird durch zwei sehr kurze Rohrabschnitte approximiert, die den Querschnitt am unteren und oberen Rand der Stimmritze repräsentieren. Für die Berechnung dieser Querschnittsflächen wird das Glottismodell von Titze [7] verwendet, welches in Abb. 3 dargestellt ist. Die Parameter des Modells sind der Abduktionsgrad,

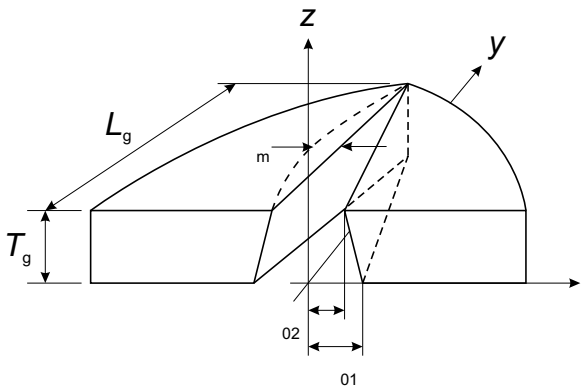

Abbildung 3: Modell der Glottis nach Titze [7].

die Phasenverschiebung zwischen dem unteren und oberen Rand der Stimmlippen und die Grundfrequenz der Schwingung. Zusätzlich wurde das Modell analog zu [3] um einen parallelen Spalt zwischen den Aryknorpeln erweitert, dessen Querschnittsfläche mit einem weiteren Parameter eingestellt werden kann. Da mit dem Glottismodell eine akustische Kopplung zwischen dem Vo-

kaltrakt und den subglottalen Luftwegen entsteht, erhält man einen sehr natürlichen Stimmton.

Das subglottale System wurden nach den Daten von Ishizaka et al. [5] modelliert und besitzt Resonanzfrequenzen bei etwa 640, 1400 und 2100 Hz, die denen eines Menschen entsprechen. Der Nasenraum wurde durch die Querschnittsfunktion von Dang et al. [4] modelliert und besitzt 4 Nasennebenhöhlen (Kreise in Abb. 1), die charakteristische Nullstellen in der Übertragungsfunktion des Nasenraums erzeugen.

Aerodynamisch-akustische Simulation

Für die aerodynamisch-akustische Simulation wird das Rohrmodell analog zu einer elektrischen Übertragungsleitung mit konzentrierten Elementen behandelt. Dazu wird jeder Rohrabschnitt durch einen Vierpol wie in Abb. 4 repräsentiert, in dem Schallfluss und Schalldruck den elektrischen Größen Strom und Spannung entsprechen. Die Netzwerkelemente und die zeitdiskrete Simulation sind in [2] erläutert. Geräuschquellen für Aspirations- und Friktionsrauschen

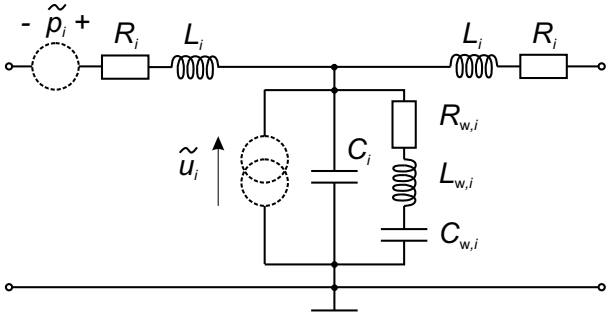

Abbildung 4: Elektrisches Schaltbild für einen Rohrabschnitt.

($\tilde{p}$ und $\tilde{u}$ in Abb. 4) werden während der Simulation unter den entsprechenden Umständen automatisch an die richtigen Positionen im Netzwerk eingefügt. Die Spektren und Amplituden der Geräuschquellen werden mit Hilfe der Strömungsgeschwindigkeit und der Querschnittsfläche in der Konstriktion sowie mit dem Abstand zwischen der Konstriktion und einem Hindernis im turbulenten Luftstrom berechnet.

Steuerung der Artikulatoren

Für die Steuerung der Artikulation wird ein gestisch motivierter Ansatz in Anlehnung an [6] in Kombination mit einem Dominanzmodell verwendet. Dazu wird für jedes Phonem zunächst ein Satz an Vokaltraktparametern definiert. Für Vokale definieren die Parameter die Sprechtraktform im gehaltenen Zustand, und für Konsonanten die Form zum Zeitpunkt der stärksten Ausbildung der Konstriktion (im symmetrischen Schwa-Kontext). Zusätzlich wird jedem Parameter pro Laut ein Dominanzwert zugewiesen, der angibt, wie wichtig die Erreichung des Parameterwertes für den entsprechenden Laut in Hinblick auf die Koartikulation ist. Beispielsweise ist der Dominanzwert des Parameters für die Zungenhöhe für /k/ 100%, so dass der Verschluss am Gaumen in jedem Fall hergestellt wird. Für die horizontale Zungenposition im /k/ ist der Dominanzwert aber nur 25%, so dass diese Position zu 75% vom Vokalkontext beeinflusst wird. Die konkreten Dominanzwerte sind bisher nur Schätzwerte, die später genauer untersucht werden sollen. Die Modellierung einer Äußerung erfolgt interaktiv durch die zeitliche Koordination überlappender Gesten, die jeweils mit einem Phonem aus der Liste assoziiert sind. Neben den Gesten für die supraglottale Artikulation müssen zusätzliche Steuerkommandos für die glottale Aktivität incl. der Grundfrequenz gegeben werden.

Schlussfolgerungen

Gegenwärtig untersuchen wir anhand von VCV-Logatomen, einzelnen Wörtern und kurzen Sätzen die zeitliche Koordination der Artikulation für eine möglichst gute Synthesequalität. Dabei hat sich gezeigt, dass mit der artikulatorischen Synthese eine hohe Sprachqualität erreicht werden kann, die mit anderen Syntheseverfahren (konkatenative Synthese, Formantsynthese) durchaus vergleichbar ist. In Zukunft streben wir eine regelbasierte Steuerung des Synthetisators auf der Basis einer phonetischen Transkription an. Eine frühere Version des Synthetisators kann im Internet unter *http://wwwicg.informatik.uni-rostock.de/~piet/speak_main.html* für Forschungs- und Ausbildungszwecke frei heruntergeladen werden.

Literatur

[1] Birkholz, P. und D. Jackèl: *A Three-Dimensional Model of the Vocal Tract for Speech Synthesis*. In: *Proceedings of the 15th International Congress of Phonetic Sciences*, S. 2597–2600, Barcelona, Spain, 2003.

[2] Birkholz, P. und D. Jackèl: *Influence of Temporal Discretization Schemes on Formant Frequencies and Bandwidths in Time Domain Simulations of the Vocal Tract System*. In: *Interspeech 2004-ICSLP*, Jeju, Korea, 2004.

[3] Cranen, B. und J. Schroeter: *Physiologically Motivated Modelling of the Voice Source in Articulatory Analysis/Synthesis*. Speech Communication, 19:1–19, 1996.

[4] Dang, J. und K. Honda: *Acoustic Characteristics of the Human Paranasal Sinuses Derived from Transmission Characteristic Measurement and Morphological Observation*. JASA, 100(5):3374–3383, 1996.

[5] Ishizaka, K., M. Matsudaira und T. Kaneko: *Input Acoustic-Impedance Measurement of the Subglottal System*. JASA, 60(1):190–197, 1976.

[6] Kröger, B. J.: *Ein phonetisches Modell der Sprachproduktion*. Niemeyer, Tübingen, 1998.

[7] Titze, I. R.: *Parameterization of the Glottal Area, Glottal Flow, and Vocal Fold Contact Area*. JASA, 75(2):570–580, 1984.