Modellierung des Sprechapparats als akustisches Netzwerk

Peter Birkholz

Institut für Akustik und Sprachkommunikation, Technische Universität Dresden

Zusammenfassung

Ein wichtiger Gegenstand der Stimm- und Sprechforschung ist die akustische Analyse und Simulation des Sprechapparats. In den 1960er Jahren wurde der Grundstein für das Verständnis der akustischen Sprachproduktion gelegt, indem man die Theorie elektrischer Netzwerke auf das akustische System des Sprechapparats anwandte. Die akustische Repräsentation des Sprechapparats in Form eines Netzwerks hat sich als effektive Abstraktionsebene erwiesen und zu großen Fortschritten in der akustischen Phonetik geführt. Sie ist auch die Grundlage für die hochgradig flexible artikulatorische Sprachsynthese. Dieser Beitrag gibt einen Überblick über die akustische Modellierung im artikulatorischen Synthetisator VocalTractLab, der aktuell die höchste Synthesequalität in der artikulatorischen Synthese erreicht.

1 Einleitung

Sprache wird durch die Bewegung der Artikulatoren, also der Zunge, des Unterkiefers, der Lippen, des Gaumensegels und des Kehlkopfes erzeugt. Diese Bewegungen bestimmen die zeitveränderliche Form des Vokaltrakts (des Hohlraums zwischen den Stimmlippen im Kehlkopf und der Mundöffnung) und damit seine akustischen Resonanzeigenschaften. Die akustische Anregung des Vokaltrakts erfolgt stimmhaft durch die Schwingung der Stimmlippen oder stimmlos durch Rauschquellen, die aufgrund von Luftturbulenzen im Bereich von Engstellen entstehen. Die Anregungssignale werden durch den Vokaltrakt gefiltert und als Sprachsignal von der Mund- und Nasenöffnung abgestrahlt. Insofern kann der Vokaltrakt als zeitveränderliches Filter mit einer zeitveränderlichen Übertragungsfunktion aufgefasst werden. Die genaue Beziehung zwischen der Vokaltraktform und seiner Übertragungsfunktion war lange Zeit ein Rätsel. Erst in den 1960er Jahren wurde eine akustische Theorie der Sprachproduktion entwickelt (Fant, 1960; Flanagan, 1965), mit der grundsätzlich für eine beliebige Vokaltraktform die Übertragungsfunktion des Vokaltrakts und damit das Sprachsignal berechnet werden konnte. Dies gelang durch die Anwendung der Theorie elektrischer Netzwerke und insbesondere der elektrischen Übertragungsleitung auf das Problem der Akustik des Vokaltrakts.

Die akustische Modellierung des Vokaltrakts ermöglicht die Synthese von Sprache auf Basis der Vokaltraktform. Dies ist die Grundlage für die artikulatorische Sprachsynthese, bei der der Prozess der Spracherzeugung auf artikulatorischer und akustischer Ebene vollständig simuliert wird. Im Vergleich zu den aktuell verbreitetsten Verfahren der Sprachsynthese, bei denen kurze Bausteine natürlicher Sprache (z. B. Silben oder Lautübergänge) zu neuen Äußerungen verkettet werden, bietet artikulatorische Synthese einen deutlich höheren Grad an Flexibilität bezüglich des Stimmklangs, des Sprechstils, und des Ausdrucks von Emotionen. Daher gilt artikulatorische Synthese als die ideale Form der Sprachsynthese (Shadle & Damper, 2002), auch wenn die erreichte Synthesequalität aufgrund der damit verbundenen Herausforderungen lange Zeit hinter der Qualität anderer Syntheseverfahren lag. Durch die stetige Weiterentwicklung der akustischen und artikulatorischen Modelle erscheint eine Ablösung der herkömmlichen Syntheseverfahren durch die viel leistungsfähigere artikulatorische Synthese aber mittelfristig möglich.

Dieser Artikel gibt einen kurzen Überblick über die akustische Modellierung des Sprechapparats für die artikulatorische Sprachsynthese VocalTractLab (Birkholz, 2005, 2013, 2014; www.vocaltractlab.de). Die Besonderheit an diesem System ist, dass nicht nur der eigentliche Vokaltrakt, sondern der *gesamte Sprechapparat* durch ein einheitliches akustisches Netzwerk repräsentiert wird.

2 Akustische Modellierung des Sprechapparats

2.1 Rohrmodell des Sprechapparats

Der Sprechapparat besteht aus allen an der Spracherzeugung beteiligten anatomischen Bereichen, d. h. aus der Lunge, den Bronchien, der Luftröhre, dem Kehlkopf, sowie dem Rachen-, Mund- und Nasenraum. Diese Bereiche sind luftgefüllte, miteinander verbundene Hohlräume (Abbildung 1a) und damit akustische Resonatoren. Für die akustische Modellierung machen wir folgende Annahmen:

- Der Sprechapparat wird als verzweigtes Rohrsystem betrachtet, in dem sich die Schallwellen nur entlang der Rohrachse ausbreiten, so dass ein ebenes Schallfeld entsteht. In dem für die Sprache wichtigen Frequenzbereich bis 5 kHz ist diese Näherung zulässig, da hier die Wellenlängen groß im Vergleich zum Rohrquerschnitt sind, und sich so keine stehenden Querwellen ausbilden. Die verzweigte Rohrachse ist in Abbildung 1a als gestrichelte Linie dargestellt. Hier gibt es ein Hauptrohr aus Luftröhre, Rachen- und Mundraum (Lunge wurde hier vernachlässigt), von dem ein Schallpfad in die Schluckrinne (Eingang zur Speiseröhre) und einer in den Nasenraum abzweigt. Vom Nasenraum wiederum zweigt das Rohr in die Nasennebenhöhlen ab, von denen hier nur zwei (von insgesamt 2 x 4 Nebenhöhlen) dargestellt sind.
- Die Krümmung der Rohrachse, insbesondere zwischen Rachen- und Mundraum, wird in der eindimensionalen Näherung des Vokaltrakts vernachlässigt (Sondhi, 1986).
- Das Rohrsystem wird durch eine Aneinanderreihung kurzer, zylindrischer Rohrabschnitte angenähert (siehe Abbildung 1b), die jeweils den Flächeninhalt des Querschnitts an der jeweiligen Position entlang der Rohrachse abbilden. Von der tatsächlichen Form der Querschnitte entlang der Rohrachse wird somit abstrahiert, was im betrachteten Frequenzbereich (s. o.) eine zulässige Näherung ist. Für die Länge der Rohrabschnitte sind 0,5 cm typisch, sie muss jedoch nicht für alle Abschnitte gleich sein. Die Nasennebenhöhlen werden gesondert als Helmholtz-Resonatoren betrachtet, bei der akustischen Simulation im Netzwerk aber analog zu den Rohrsegmenten behandelt (siehe unten).

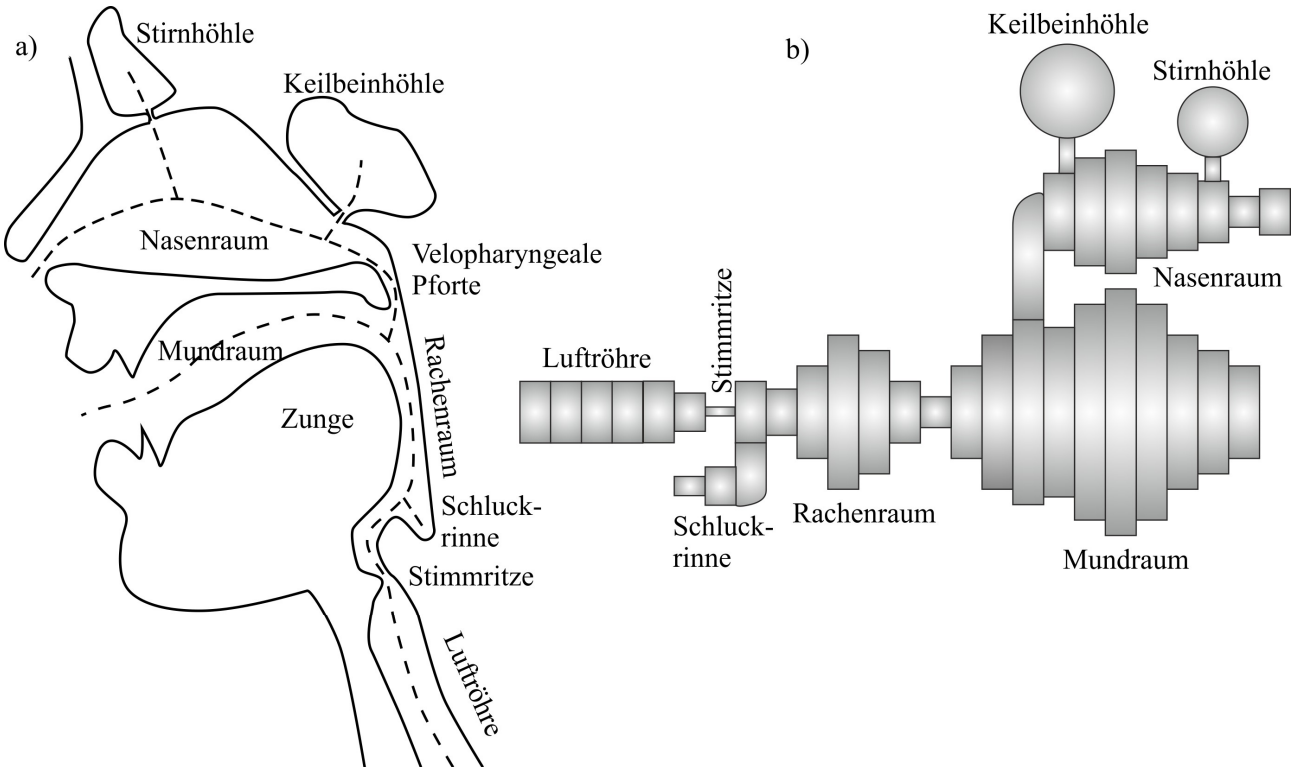

Abbildung 1 a) Darstellung der Schallpfade im Sprechapparat (gestrichelte Linien). b) Entsprechendes Rohrmodell.

2.2 Netzwerkrepräsentation von Rohrabschnitten

Mit Hilfe elektro-akustischer Analogien (Beranek, 1954; Flanagan, 1965) lässt sich ein kurzer zylindrischer Rohrabschnitt durch einen elektrischen Vierpol wie in Abbildung 2a darstellen. In dieser Analogie entspricht der Volumenstrom (Schallfluss) *u* im Rohr dem elektrischen Strom *I*, und der Schalldruck *p* der elektrischen Spannung *U*. Das Bezugspotential (Masse) entspricht dem Luftdruck außerhalb des Vokaltrakts. Die Induktivität *L* und die Kapazität *C* des Vierpols werden im Folgenden beispielhaft aus der Bewegungsgleichung und der Kontinuitätsgleichung des eindimensionalen Schallfeldes abgeleitet.

Die *Bewegungsgleichung* beschreibt die Änderung der Schallschnelle aufgrund eines Druckgradienten, und kann für den 1D-Fall als

$$\frac{\partial p}{\partial x} = -\rho_0 \frac{\partial v}{\partial t}$$

geschrieben werden, wobei *x* die Ortskoordinate, *t* die Zeit, $p = p(x,t)$ der Schalldruck, $v = v(x,t)$ die Schallschnelle, und ρ_0 die mittlere Dichte der Luft ist. Wenn die Schallwelle einen Rohrabschnitt mit dem konstanten Querschnitt *A* durchläuft, so lässt sich die Gleichung auch mit Hilfe des Volumenstroms $u(x,t) = A \cdot v(x,t)$ schreiben:

$$\frac{\partial p}{\partial x} = -\frac{\rho_0}{A} \frac{\partial u}{\partial t}$$

Für ein hinreichend kurzes ortsfestes Rohrstück der Länge l lässt sich die linke Seite durch einen Differenzenquotienten approximieren, womit u nur noch eine Funktion der Zeit wird:

$$\frac{p_2 - p_1}{l} = -\frac{\rho_0}{A} \frac{du}{dt}$$

bzw.

$$p_1 - p_2 = \frac{\rho_0 l}{A} \frac{du}{dt} = L \frac{du}{dt}$$

Diese Gleichung ist formal identisch zur Strom-Spannungs-Beziehung an einer Spule, wobei die Schalldruckdifferenz $p_1 - p_2$ der Spannungsdifferenz über der Spule und u dem Strom durch die Spule entspricht. Die analoge akustische Induktivität („akustische Masse“) ist demnach $L = (\rho_0 l)/A$. Der Vierpol zu dem Rohrabschnitt besitzt eine T-Struktur, wobei man die Induktivität gleichmäßig auf zwei Spulen im linken und rechten Zweig aufteilt, so dass für Abbildung 2a gilt: $L = \rho_0 l/(2A)$.

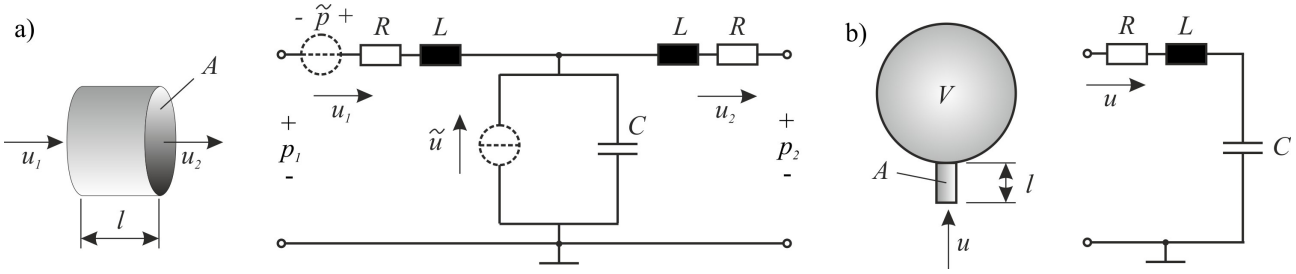

Abbildung 2 a) Vierpolnetzwerk eines kurzen Rohrabschnitts. b) Zweipolnetzwerk eines Helmholtz-Resonators.

Die *Kontinuitätsgleichung* des eindimensionalen Schallfeldes beschreibt die Änderung des Schalldrucks aufgrund eines Gradienten der Schallschnelle *v*:

$$\frac{\partial v}{\partial x} = -\frac{1}{\rho_0 c^2} \frac{\partial p}{\partial t}$$

Hierbei ist c die Schallgeschwindigkeit. Mit der Ersetzung $v(x,t) = u(x,t)/A$ erhalten wir wiederum

$$\frac{\partial u}{\partial x} = -\frac{A}{\rho_0 c^2} \frac{\partial p}{\partial t}.$$

Für einen hinreichend kurzen Rohrabschnitt der Länge l können wir die linke Seite durch den Differenzenquotienten $(u_2 - u_1)/l$ annähern, wobei u_1 der Volumenstrom am Rohreingang und u_2 der Volumenstrom am Rohrausgang sind. Damit erhalten wir nach Umstellung

$$u_1 - u_2 = \frac{Al}{\rho_0 c^2}\frac{dp}{dt} = C\frac{dp}{dt}.$$

Diese Gleichung ist formal identisch zur Strom-Spannungs-Beziehung an einem Kondensator, wobei $u_1 - u_2$ dem in den Kondensator fließenden Strom entspricht und p der Spannung über dem Kondensator. $C = Al/(\rho_0 c^2)$ ist die „akustische Federung" des Rohrabschnitts und repräsentiert die Kompressibilität der Luft.

Wären im Vierpol in Abbildung 2a nur die zwei Spulen mit der Induktivität L und der Kondensator mit der Kapazität C vorhanden, so wäre dies die Netzwerkrepräsentation des Rohrabschnitts für den verlustfreien Fall. In der Realität ist die Schallausbreitung in einem Rohr aufgrund der Reibung jedoch stets verlustbehaftet. Dies lässt sich durch den Längswiderstand *R* in dem Vierpol darstellen. Der Widerstandswert ist nicht direkt aus den 1D-Schallfeldgleichungen ableitbar und ist außerdem frequenzabhängig. In Hinblick auf eine zeitdiskrete numerische Simulation des Netzwerks ist es sinnvoll, den tatsächlichen Widerstand durch eine frequenzunabhängige Näherung zu beschreiben, z. B. durch den Widerstand einer laminaren stationären Strömung durch ein zylindrisches Rohr mit dem Querschnitt A. Für das R in Abbildung 2a ergibt sich dann (Birkholz, 2005)

$$R = \frac{8\mu\pi}{A^2}\frac{l}{2},$$

wobei μ die dynamische Viskosität der Luft ist.

Der Vollständigkeit halber sind in Abbildung 2a gestrichelt noch potenzielle Schalldruck- oder Volumenstromquellen eingezeichnet. Diese Quellen können verwendet werden, um im Bereich turbulenter Strömung (die mit dem 1D-Modell selbst nicht modelliert werden kann) Rauschen in den Vokaltrakt einzuspeisen. Die Einfügung solcher Rauschquellen ist für die Simulation von Reibe- und Zischlauten wichtig und wird z. B. in Birkholz (2014) diskutiert.

Die Nasennebenhöhlen werden akustisch effektiv durch Helmholtz-Resonatoren modelliert (Dang et al., 1994). Ein Helmholtz-Resonator besteht aus einem kugelförmigen Hohlraum mit dem Volumen *V*, der mit einem kurzen Rohrstück (Hals) mit dem Querschnitt *A* und der Länge *l* verbunden ist. Abbildung 2b zeigt das analoge Zweipolnetzwerk für einen solchen Resonator. *L* und *R* sind die akustische Masse und der akustische Widerstand des Halses und werden analog zum „normalen" Rohrabschnitt berechnet. *C* ist die akustische Federung des Kugelhohlraums und wird als $C = V/(\rho_0 c^2)$ berechnet.

2.3 Netzwerk für den gesamten Sprechapparat

Wenn die Netzwerkabschnitte aller Rohrabschnitte des verzweigten Rohrmodells miteinander verkettet werden, so erhält man das in Abbildung 3 dargestellte Netzwerk für den gesamten Sprechapparat. Dieses Netzwerk entspricht einer verzweigten inhomogenen elektrischen Übertragungsleitung mit diskreten Elementen. Der erste Rohrabschnitt der Luftröhre wird am Eingang mit einer Druckquelle (analog zur elektrischen Spannungsquelle) verbunden, die den Überdruck in der Lunge gegenüber dem atmosphärischen Druck nachbildet. An den vordersten Rohrabschnitten (Vierpolen) des Nasenraums und des Mundraums wird das Netzwerk jeweils mit einer Strahlungsimpedanz abgeschlossen. Die Strahlungsimpedanz kann bis ca. 5 kHz in guter Näherung als eine Parallelschaltung aus einem akustischen Widerstand R_{rad} und einer akustischen Masse L_{rad} modelliert werden (Flanagan, 1965), wobei diese von der abstrahlenden Fläche A_{rad}, also der Mund- oder Nasenöffnung abhängen:

$$R_{rad} = \frac{128\rho_0 c}{9\pi^2 A_{rad}}$$

und

$$L_{rad} = \frac{8\rho_0}{3\pi\sqrt{A_{rad}\pi}}.$$

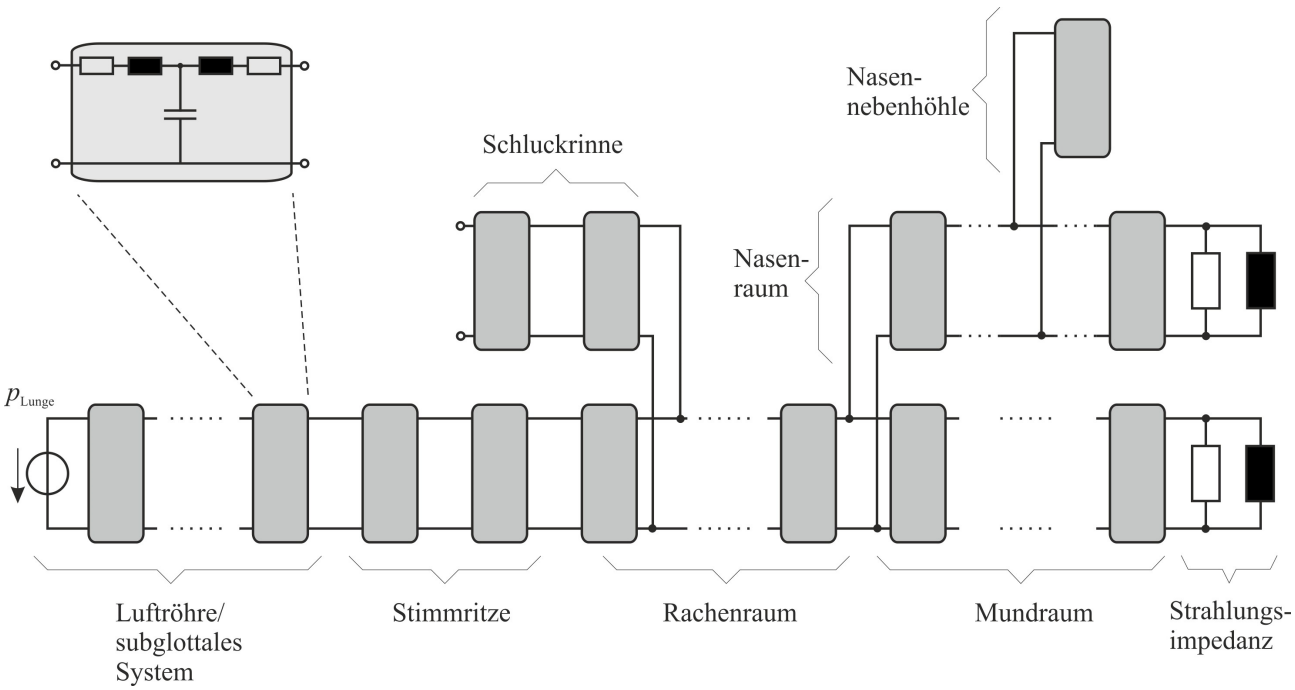

Abbildung 3 Akustisches Netzwerk für den gesamten Sprechapparat.

3 Synthese von Sprache

Das oben entwickelte Netzwerk ist ein anschauliches und hinreichend genaues Modell des akustischen Verhaltens des Sprechapparats. Die Netzwerkelemente sind dabei unmittelbar von der Vokaltraktform und damit von der Rohrgeometrie abhängig, die sich beim Sprechen kontinuierlich verändert. Um das Netzwerkmodell für die artikulatorische Sprachsynthese einzusetzen, werden daher ein Vokaltraktmodell und ein Verfahren zu dessen Ansteuerung benötigt, die zu jedem Zeitpunkt die Rohrgeometrie definieren (z. B. Birkholz, 2013; Engwall, 2003; Mermelstein, 1973). Üblicherweise wird nur der Rachen- und Mundraum artikulatorisch modelliert, da nur dieser seine Form beim Sprechen wesentlich verändert. Die Luftröhre (und ggfs. die Bronchien und Lunge) sowie der Nasenraum werden dagegen als quasi-statisch betrachtet, und durch zeitlich unveränderliche Rohrabschnitte repräsentiert (Weibel, 1963, Dang et al., 1994). Die Stimmlippen, und damit die Rohrabschnitte, die die Stimmritze repräsentieren, werden wiederum separat modelliert, da sich die Stimmlippen im Vergleich zu anderen Artikulatoren wie Zunge oder Lippen sehr schnell bewegen. Die schnelle Bewegung der Stimmlippen verursacht ein quasi-periodisches Öffnen und Schließen der Stimmritze und erzeugt damit den Stimmton beim Sprechen. Einen Überblick über selbstschwingende biomechanische Modelle der Stimmlippen geben Birkholz (2011) und Erath et al. (2013).

Um das eigentliche Sprachsignal auf Basis des zeitveränderlichen Netzwerkmodells zu berechnen, muss das Schallfeld simuliert werden. Hier kommen numerische Methoden zum Einsatz, die zu jedem Zeitpunkt der Simulation die Schalldrücke in allen Knoten und die Volumenströme in allen Zweigen des Netzwerks berechnen (z. B. Birkholz & Jackèl, 2004; Maeda, 1982). Da die Drücke und Volumenströme alle voneinander abhängig sind, läuft diese Berechnung auf die Lösung eines großen linearen Gleichungssystems mit einer dünn besetzten Koeffizientenmatrix hinaus. Eine typische Abtastrate für solche Simulation ist 44100 Hz. Das eigentliche Sprachsignal ist der von der Mundöffnung abgestrahlte Schalldruck. Dieser kann durch die erste zeitliche Ableitung des Volumenstroms durch die Strahlungsimpedanzen an der Mund- und Nasenöffnung angenähert werden (Stevens, 2000).

Das in diesem Beitrag skizzierte Prinzip ist die Grundlage für die artikulatorische Sprachsynthese VocalTractLab, die mit dem Ziel entwickelt wird, synthetische Sprache höchster Natürlichkeit bei gleichzeitig größter Flexibilität in Bezug auf die Synthesestimme zu erzeugen. Audiobeispiele der Synthese sind unter www.vocaltractlab.de zu finden. Abbildung 4 zeigt beispielhaft das Oszillogramm und Spektrogramm der

Äußerung „Die Tür ist offen“ eines echten Sprechers und im Vergleich dazu das Oszillogramm und Spektrogramm der mit VocalTractLab nachsynthetisierten Äußerung. Hierbei ist eine gute Übereinstimmung festzustellen, die das Potenzial artikulatorischer Synthese aufzeigt.

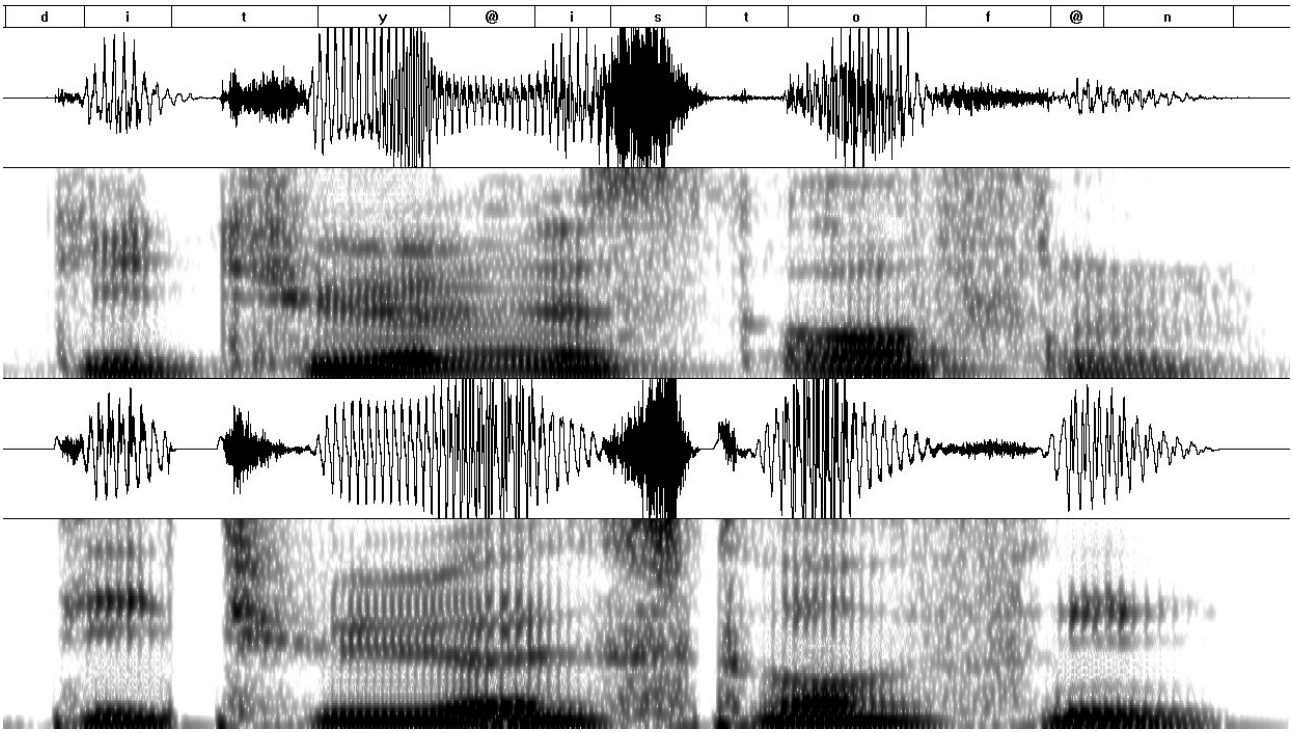

Abbildung 4 Oben: Oszillogramm und Spektrogramm einer natürlichsprachlichen Aufnahme des Satzes „Die Tür ist offen“. Unten: Oszillogramm und Spektrogramm der artikulatorischen Resynthese des obigen Satzes.

Literatur

L. L. Beranek: Acoustics. New York: McGraw-Hill 1954.

P. Birkholz: 3D-Artikulatorische Sprachsynthese. Berlin: Logos Verlag 2005.

P. Birkholz: A survey of self-oscillating lumped-element models of the vocal folds. In: B. J. Kröger, P. Birkholz (Hrsg.): Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung, Dresden TUDPress 2011, S. 47-58.

P. Birkholz: Modeling consonant-vowel coarticulation for articulatory speech synthesis. PLoS ONE 8 (2013), e60603. doi:10.1371/journal.pone.0060603.

P. Birkholz: Enhanced area functions for noise source modeling in the vocal tract. In: Proc. of the 10th International Seminar on Speech Production (ISSP 2014), Köln 2014, S. 37-40.

P. Birkholz, D. Jackèl: Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system. In: Proc. of the Interspeech 2004-ICSLP, Jeju, Korea 2004, S. 1125–1128.

P. Birkholz, B. J. Kröger, C. Neuschaefer-Rube: Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis. In: Proc. of the Interspeech 2011, Florenz 2011, S. 2681-2684.

J. Dang, K. Honda, H. Suzuki: Morphological and acoustical analysis of the nasal and the paranasal cavities. The Journal of the Acoustical Society of America 96 (1994), S. 2088-2100.

O. Engwall: Combining MRI, EMA and EPG measurements in a three-dimensional tongue model. Speech Communication 41 (2003), S. 303-329.

B. D. Erath, M. Zañartu, K. C. Stewart, M. W. Plesniak, D. E. Sommer, S. D. Peterson: A review of lumped-element models of voiced speech. Speech Communication 55 (2013), S. 667-690.

G. Fant. Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations. Walter de Gruyter 1960.

J. L. Flanagan: Speech analysis, synthesis and perception. Berlin: Springer-Verlag 1965.
S. Maeda: A digital simulation method of the vocal-tract system. Speech Communication 1 (1982), S. 199-229.
P. Mermelstein: Articulatory model for the study of speech production. The Journal of the Acoustical Society of America 53 (1973), S. 1070-1082.
C. H. Shadle, R. I. Damper: Prospects for articulatory synthesis: A position paper. In: Proc. of 4th ISCA Workshop on Speech Synthesis. Pitlochry, Scotland 2002, S. 121–126.
M. M. Sondhi: Resonances of a bent vocal tract. The Journal of the Acoustical Society of America 79 (1986), S. 1113-1116.
K. N. Stevens: Acoustic phonetics. Cambridge: MIT Press 2000.
E. R. Weibel: Geometry and dimensions of airways of conductive and transitory zones. Berlin: Springer-Verlag 1963.