Artikulatorisch-akustische Simulation der Spracherzeugung

Peter Birkholz

Klinik für Phoniatrie, Pädaudiologie und Kommunikationsstörungen, Universitätsklinikum Aachen und RWTH Aachen University

Dieser Beitrag stellt das Computersystem VocalTractLab für die artikulatorische und akustische Simulation der Spracherzeugung auf Basis eines dreidimensionalen Vokaltraktmodells vor. Es ermöglicht die Synthese beliebiger Sprachäußerungen in hoher Qualität und stellt dabei die artikulatorischen und akustischen Mechanismen beim Sprechen nachvollziehbar durch Grafiken und Animationen dar. Der Kern des Systems ist ein dreidimensionales geometrisches Modell des Vokaltrakts (Birkholz, 2013). Die Modellartikulatoren (Zunge, weicher Gaumen, Lippen, etc.) werden mit einer Reihe von Parametern gesteuert, wodurch das Modell alle wesentlichen Vokaltraktformen annehmen kann, die bei der Realisierung von Sprachlauten auftreten. Die akustische Anregung des Vokaltrakts wird auf Basis eines biomechanischen, selbstschwingenden Modells der Stimmlippen simuliert, welches auf dem klassischen Zwei-Massen-Modell von Ishizaka & Flanagan (1972) basiert. Dieses Modell ist in der Lage, die wesentlichen Eigenschaften realer Stimmlippen bei der Phonation nachzubilden. Insbesondere haben wir es kürzlich dahingehend erweitert, in realistischer Weise unterschiedliche Phonationstypen bzw. Stimmqualitäten zu simulieren, insbesondere das Kontinuum von gepresster über modale zu behauchter Stimme (Birkholz et al., 2011). Die Stimmqualität wird während des Sprechens ständig variiert (z. B. zwischen betonten und unbetonten Silben), und ist besonders für den Ausdruck von Emotionen oder stimmlichem Affekt ein wichtiges paralinguistisches Merkmal. Neben den Modellen für den Vokaltrakt und die Stimmlippen wurden auch das subglottale System und der Nasenraum modelliert. Gemeinsam werden die Parameter der Modelle durch eine „gestische Partitur“ gesteuert, in der artikulatorische Gesten zeitlich miteinander koordiniert werden müssen, um eine bestimmte Äußerung zu erzeugen (Birkholz, 2007a).

Für die aerodynamisch-akustische Simulation des Sprechens werden alle diese Modelle durch ein einheitliches verzweigtes Rohrsystem des Sprechapparats repräsentiert, in dem mittels numerischer Simulation im Zeitbereich die relevanten Schall- und Strömungsgrößen berechnet werden (Birkholz, 2005; Birkholz et al., 2007). Neben der Berechnung des abgestrahlten Schalldrucksignals erlaubt dies auch die detaillierte zeitliche Analyse von Schall- und Strömungsgrößen an beliebigen Stellen innerhalb des Sprechapparats.

Das Simulationssystem wird bereits von zahlreichen Instituten in der Lehre und für Forschungszwecke eingesetzt. Unser Ziel ist es vorrangig, die Simulation dahingehend zu verbessern, dass die synthetische Sprache hinsichtlich der Natürlichkeit nicht oder kaum noch von natürlicher Sprache zu unterscheiden ist. Dies würde dann auch Anwendungen in der Text-zu-Sprach-Synthese ermöglichen, wo derzeit Systeme dominieren, die auf der Verkettung voraufgenommener fixer Sprachbausteine basieren. Die hier beschriebene „artikulatorische Sprachsynthese“ wäre jedoch deutlich flexibler bezüglich der Simulation unterschiedlicher Sprecher, Sprechstile und Emotionen als es derzeitige Verfahren zur Sprachsynthese sind. Weitere Anwendungen unseres Systems sind z. B. psychologisch-perzeptive Untersuchungen (z. B. Xu et al., 2013), die Synthese von Gesang (Birkholz, 2007b) und die neuronale Simulation des Spracherwerbs (Kröger et al., 2009).

Referenzen

Birkholz P (2005). 3D-Artikulatorische Sprachsynthese. Logos Verlag, Berlin

Birkholz P (2007a). Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets. In Proc. of the Interspeech 2007 - Eurospeech, pp. 2865–2868, Antwerp, Belgium

Birkholz P (2007b). Articulatory synthesis of singing. In: Bloothooft G (ed.) Synthesis of Singing Challenge. Antwerp, Belgium, URL: http://www.let.uu.nl/~Gerrit.Bloothooft/personal/SSC/ index.htm

Birkholz P, Jackèl D, Kröger BJ (2007). Simulation of losses due to turbulence in the time-varying vocal system. IEEE Transactions on Audio, Speech and Language Processing, 15(4), pp. 1218–1226

Birkholz P, Kröger BJ, Neuschaefer-Rube C (2011). Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis. In Proc. of the Interspeech 2011, pp. 2681-2684, Florence, Italy

Birkholz P (2013). Modeling consonant-vowel coarticulation for articulatory speech synthesis. PLoS ONE, 8(4): e60603. doi:10.1371/journal.pone.0060603

Ishizaka K, Flanagan JL (1972). Synthesis of voiced sounds from a two-mass model of the vocal cords. The Bell System Technical Journal 51(6), 1233–1268

Kröger BJ, Kannampuzha J, Neuschaefer-Rube C (2009) Towards a neurocomputational model of speech production and perception. Speech Communication 51: 793-809

Xu Y, Lee A, Wu W-L, Liu X, Birkholz P (2013). Human vocal attractiveness as signaled by body size projection. PLoS ONE, 8(4): e62397. doi:10.1371/journal.pone.0062397