Faszination Sprechende Maschinen: Technologischer Wandel der Sprachsynthese über zwei Jahrhunderte

Lars Engeln[1], Rainer Groh[1], Falk Gabriel[1], Peter Birkholz[1], Rainer Jäckel[1], Rüdiger Hoffmann[1], Judith Felten[1], Regina Bergmann[1], Joachim Scharloth[1], Lisa Lüneburg[1], Jens Krzywinski[1], Jörg Neumann[2] und Peter Plaßmeyer[2]

[1] *TU Dresden, 01062 Dresden, Deutschland, Email: lars.engeln@tu-dresden.de, ruediger.hoffmann@tu-dresden.de*

[2] *Staatliche Kunstsammlungen Dresden, Mathematisch-Physikalischer Salon, 01067 Dresden, Deutschland, Email: peter.plassmeyer@skd.museum*

Einleitung

Die Technische Universität Dresden besitzt mit der historischen akustisch-phonetischen Sammlung (HAPS) eine Sammlung von Apparaten, Maschinen und Gegenständen, die die Entwicklung der Experimentalphonetik und Sprachtechnologie von der Mitte des 18. Jh. bis in die zweite Hälfte des 20. Jh. in einer für Europa einmaligen Geschlossenheit darstellt [7]. Die Bestände reichen von Repliken der ersten mechanischen Vokalresonatoren und Sprechapparate von Kratzenstein (1781) und von Kempelen (1791) über diverse mechanische Apparate zur Erfassung sprechphysiologischer Parameter (z. B. des Stimmtons oder der Lippen- und Kehlkopfbewegung) bis hin zu den ersten analogen elektrischen Sprachanalyse- und Synthesemaschinen (z. B. Spektrographen und Formantsynthetisatoren) (vgl. [8, 10]).

Einen großen Teil der Sammlung bilden frühe Apparate zur künstlichen Erzeugung von Sprache, welche seit jeher zur Inklusion von stummen und blinden Menschen dienen sollten. Verfahren zur Sprachsynthese gewinnen in jüngster Zeit zunehmend an Bedeutung, da sie wesentlicher Bestandteil von sprachgesteuerten Dialog- und Assistenzsystemen sind (z. B. von Navigationssystemen oder *persönlichen Assistenten* wie Alexa, Cortana und Siri). Außerdem besitzt die Sprachsynthese großes Potential für die Unterstützung des Erlernens von Fremdsprachen und deren Aussprache sowie bei der automatischen Übersetzung und hilft die sprachlichen Herausforderungen der Globalisierung zu unterstützen. Das Projekt *Faszination Sprechende Maschinen* zielt darauf ab, einen nachhaltigen Beitrag zur Dokumentation der historischen Entwicklung der Sprachsynthese im Kontext der Mensch-Technik-Interaktion zu leisten.

Im Rahmen des Projektes wurde die HAPS weiter erschlossen und es ist eine digitale Ausstellung entstanden, die ausgewählte Objekte für Studierende, Lehrende und Forschende zugänglich macht. Weiterhin ist eine moderne und modulare Sprechmaschine entstanden, die zur aktuellen Erforschung der artikulativen Sprachsynthese dient.

Erschließung der HAPS

Die Katalogisierung der HAPS begann bereits 2012 vorrangig mit der Erfassung von experimentalphonetischen Geräten [10]. Mit Projektbeginn wurden die fortführenden Katalogisierungsarbeiten auf die für die Geschichte der Sprachsynthese relevanten Bestände konzentriert. Dabei wurden die folgenden Bestandsgruppen vollständig erfasst:

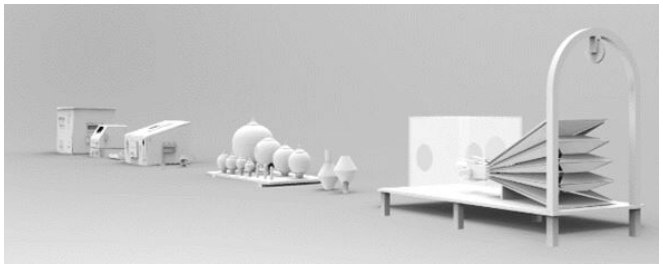

Abbildung 1: CAD-Nachbauten und 3D-Scans ausgewählter Exponate zur Entwicklung der Sprachsynthese.

- Objekte (hauptsächlich Repliken) aus der Entwicklung der mechanischen und elektromechanischen Sprachsynthese (13 Objekte)
- Objekte aus der Geschichte der elektronischen Sprachsynthese (53 Objekte)
- Objekte zur Dokumentation der Anwendung in Spielzeugen und Lernsystemen (43 Objekte)

Zusätzlich wurden die folgenden Bestandsgruppen vollständig bearbeitet, die für das Projekt inhaltlich flankierende Bedeutung haben:

- Anatomische Modelle und Lehrmittel, hauptsächlich zur menschlichen Spracherzeugung als Vorbild für die Sprachsynthese (26 Objekte)
- Objekte aus der Geschichte der elektronischen Analyse und Erkennung von Sprache (31 Objekte)

Insgesamt sind nun 423 Objekte der Sammlung erfasst. Für die Virtuelle Ausstellung wurden 12 Objekte als Repräsentaten der Entwicklung der Sprachsynthese ausgewählt und digitalisiert (siehe Abb. 1). Zunächst wurden die Objekte direkt am Aufbewahrungsort der Sammlung gescannt. Diese 3D-Modelle mussten dann bereinigt und für die interaktive Nutzung aufbereitet werden. Zum Teil wurden glänzende, rotationssymmetrische Objekte wegen starker Degenerierung im Scan nachmodelliert (CAD).

Eine besondere Vertiefung der Untersuchungen ergibt sich aus dem Umstand, dass sich in der HAPS eine

Kollektion mechanischer Stimmen befindet, von denen bekannt ist, dass 1899 ihre Anwendung beim Training hörbehinderter Patienten von dem Jenaer Otologen Johannes Kessel (1839 – 1907) vorgeschlagen wurde. Kurioserweise ist der Phonetikgeschichte nie aufgefallen, dass es sich bei den mechanischen Stimmen um Adaptionen aus der Spielzeug- und Automatenindustrie handelt, deren Ursprung sich bis auf die Sprechmaschine von W. von Kempelen zurückführen lässt [9].

Für den erweiterten Kontext wurde eine umfangreiche Literatur- und Patentrecherche zur Entwicklung der Sprachsynthese in der früheren UdSSR realisiert [5, 6]. Ausgewertet wurden Monographien, Zeitschriftenartikel, Konferenzbeiträge und Patentschriften führender Vertreter der sowjetischen Akustik und Sprachsignalverarbeitung. Bis in die jüngste Zeit wurde die Entwicklung der Sprachtechnologie in der früheren UdSSR fast ausschließlich aus literarischen Quellen rezipiert, da die wichtigsten Dokumente bis um die Jahrtausendwende unter Verschluss gehalten wurden. K. F. Kalachev (1915 – 2001), der bereits in der Vorkriegszeit gemeinsam mit V. A. Kotel'nikov (1908 – 2005) an Problemen der verschlüsselten Sprachübertragung gearbeitet hatte, war in dem geheimen nachrichtentechnischen Labor des KGB in Marfino als Arbeitsgruppenleiter für die Entwicklung der Chiffriereinheit des ersten sowjetischen Vocoders verantwortlich gewesen. Dies zeigt, dass insbesondere die Sprachtechnologieforschung der USA und der UdSSR bis zur Einführung der Halbleiter parallel und vor allem gleichauf abliefen.

Virtuelle Ausstellung

Unter dem übergreifenden Erzählstrang *Geschichte der Sprachsynthese* wurden die eigens erstellten 3D-Modelle der ausgewählten Sammlungsobjekte textuell wie visuell mit Informationen angereichert. Dabei wurde insbesondere der multidisziplinäre Entstehungskontext der einzelnen Objekte berücksichtigt. Die Objekte wurden hinsichtlich des linguistischen und phonetischen Wissens, das sich in den Objekten niederschlägt, aber auch in Bezug auf das wissenschaftliche Arbeiten und den historisch-kulturellen Kontext zur Entstehungszeit betrachtet. Es gehört zu der besonderen Leistung der Ausstellung, dass sie eine umfangreiche Auseinandersetzung mit der Geschichte der Sprachsynthese bietet, die wiederum weitere verschiedene Disziplinen anspricht und diese auch bereichert. Durch die Expansion in den digitalen Raum wird die Sammlung sichtbarer und zugänglicher. Gleichzeitig können Studierende und Lehrende die Ausstellung für ihren Lernprozess bzw. ihre Seminargestaltung nutzen, da insbesondere die visuellen Anreicherungen den Gegenstand der Sprachsynthese *greifbar* machen.

Neben der diachronen Anordnung werden die Objekte ebenfalls synchron betrachtet. Dafür wurden die einzelnen Themenschwerpunkte Linguistik, Phonetik, Wissenschaft und der historisch-kulturelle Kontext sowie die Biographie der Entwickler der einzelnen Objekte als *Kontextblasen* um das Objekt herum angeordnet. Die zahlreichen Wissensströmungen, die die Entstehung der einzelnen Objekte bedingt haben, erscheinen so komprimierter und regen zur vertieften Auseinandersetzung an. Gleichzeitig können die Nutzer und Nutzerinnen hier gezielt auf die Themengebiete zugreifen, die sie interessieren.

Ein Ansatz zur Vermittlung diachroner Kontextsichten ist ein zeitorientierter, rotierbarer, semantischer Strahl (siehe Abb. 2, *oben links*). Um den Strahl herum liegt der Kontext mit dessen Handlungssträngen. Durch die Rotation um den Strahl werden die Handlungsstränge exploriert. Ein Exponat wird so aus verschiedenen Perspektiven beleuchtet, um ein ganzheitliches Verständnis zu erzeugen. Zu einem Exponat können alle Kontextsichten dargestellt werden (Querschnitt des Strahls). Innerhalb des Strahls werden die Objekte angezeigt. Dies dient als Überblick. Die mit einer Ontologie modulierten Daten der Ansichten können je nach Exponat historische bzw. aktuelle Bilder, Informationsgrafiken, Klangbeispiele und andere Medienobjekte bis hin zu interaktiven 3D-Modellen sein [3]. Der kontextualisierte Strahl stellt insgesamt ein zoomable User Interface dar. Der Fokus liegt somit auf einer explorativen Wissensvermittlung. Da keine Virtuelle Sammlung mit allen Objekten, sondern eine Ausstellung mit ausgewählten Exponaten erstellt wurde, beschränkt sich das Suchinterface auf Zoom und *details-on-demand* der Objekte sowie Erzählstränge. Über agile Software- und Designmethoden wurde der Demonstrator in vielen Iterationen, verfeinert oder abgeändert. Insbesondere durch die Beteiligung vieler verschiedener Fachdomänen haben diese Methoden den Diskurs gefördert. Die Ergebnisse der Iterationen wurden in Expertengesprächen evaluiert und hinterfragt. Für den Erhalt der Accessibility wird die gerenderte 3D-Szene so mit 2D-Annotationen verknüpft, dass die Texte weiterhin für Screenreader zugreifbar sind und dennoch ein interaktives Nutzerlebnis entsteht. Beim klassischen 3D-Rendering wird Text in eine Rastergrafik überführt; das heißt, dass der Text (nun ein Bild) als solcher syntaktisch verloren geht. Der Text kann nicht mehr kopiert oder von Screenreadern gelesen werden. Daher wird die 3D-Szene zunächst gerendert und daran dann passend gestyltes HTML angegliedert. So werden mehrere Schichten aufeinander komponiert.

Zur möglichen Erweiterung der Vermittlung von Audiosignalen besteht Potential in der Virtuellen Realität, um spielerisch Signalverarbeitung verständlich zu machen [4].

Moderne Sprechmaschine

W. v. Kempelens sprechende Maschine ist im funktionalen Sinne anthropomorph: der Luftstrom wird durch einen Blasebalg erzeugt, die Stimmmechanik mit aufschlagender Zunge aus Elfenbein imitiert die Schwingung der Stimmlippen, eine Windlade sorgt für den Druckausgleich, der Vokaltrakt oberhalb der Glottis besteht aus einem Mundtrichter und Nasenöffnungen. Zur Hervorbringung unterschiedlicher Vokale musste der Mundtrichter mit der linken Hand verformt beziehungsweise in bestimmter Weise abgedeckt werden. W. v. Kempelens sprechende Maschine ist oft nachgebaut worden, u.a.

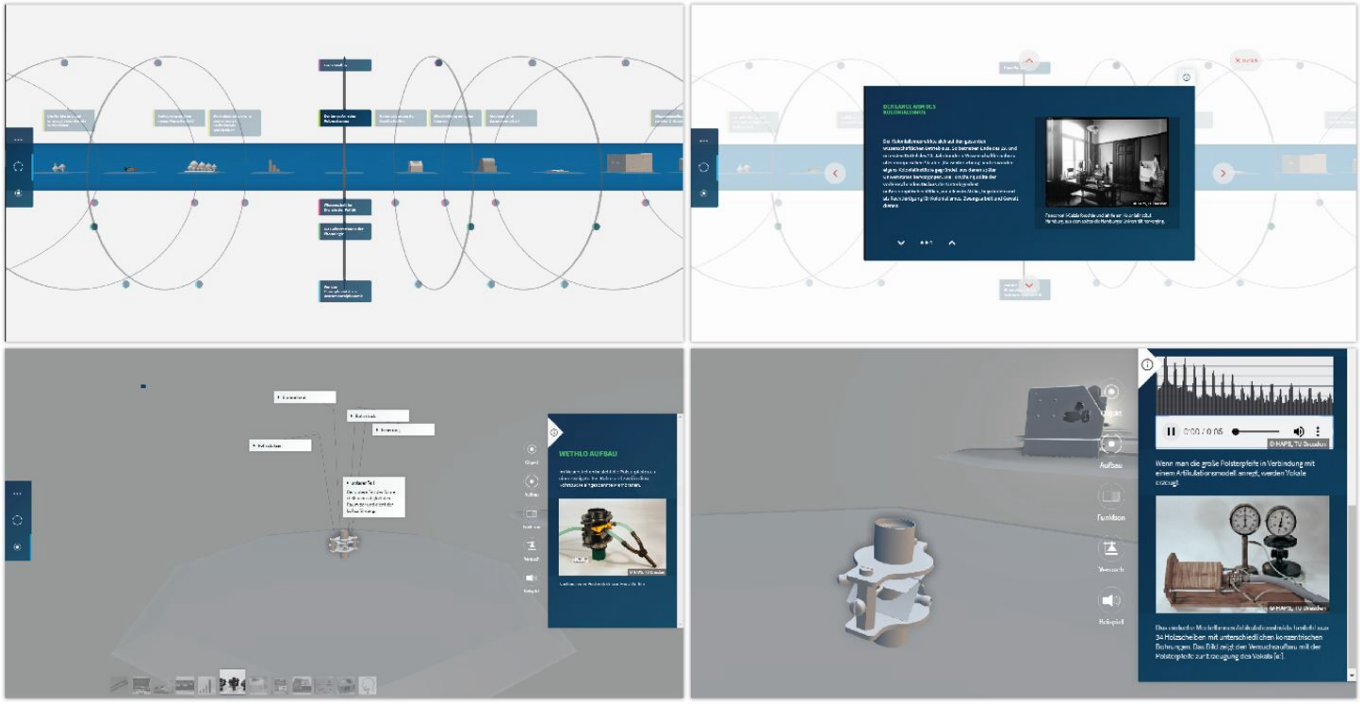

Abbildung 2: Virtuelle Ausstellung: semantischer Strahl mit *Kontextringen* (synchron vertical, diachron horizontal) *(oben links)*, aufgeklappte *Kontextblase (oben rechts)*, Objektebene mit Annotationen *(unten links)* und Beispielen *(unten rechts)*.

durch den englischen Physiker und Erfinder Charles Wheatstone (1802 – 1875), der die Apparatur für physikalisch-akustische Experimente benutzte.

Für die aktuelle Erforschung artikulatorischer Sprachsynthese wurde ein neuartiger anthropomorpher Sprechapparat mit modernen Materialien entwickelt. Es sollen Abstraktionsgrade des menschlichen Vokaltraktes untersucht werden, um diese z.B. in Modelle zur digitalen Synthese zu überführen. Dafür wurde ein menschzentriertes Konzept zum Kommunizieren und Erlebbarmachen von Funktionsweisen und Eigenschaften des Vokaltrakts entwickelt, auf dessen Basis Prinzipvarianten entsprechender Demonstratoren entworfen wurden. Danach ist ein CAD-Modell zur Fertigung (siehe Abb. 3) ausgearbeitet worden. Ergebnis daraus ist ein Vokaltraktdemonstrator, der zum einen der Veranschaulichung der Abläufe innerhalb des Sprachtrakts in der Lehre dienen kann und zum anderen eine Forschungsplattform darstellt. Das bedeutet, dass auch nach Fertigstellung des Demonstrators einzelne Elemente modular, je nach Erkenntnisstand, ausgetauscht und weiterentwickelt werden können. Dabei wurde darauf geachtet, dass der Gesamteindruck des Demonstrators mit einem menschlichen Oberkörper assoziiert werden kann. Dadurch soll ein intuitives Verständnis von Studierenden oder Fachfremden für das Thema Sprachproduktion gefördert werden. Es wurde als erster Schritt ein künstliches Lungenvolumen detailkonstruiert und durch 3D-Druck gefertigt (siehe Abb. 4). Das Lungenvolumen ist zugleich die tragende Basis des Demonstrators und überträgt die Kräfte über vier Standfüße in die Standfläche. Die Hauptfunktion des Lunge ist es, ein Ausgleichsvolumen für den zugeführten Luftstrom zu bilden. Alle weiteren Elemente des Demonstrators bauen auf diesem Volumen auf.

Die Nachahmung realer Stimmlippen wurde durch synthetische Modelle mit mehreren (Silikon-)Schichten angestrebt. Für ein realitätsnahes Glottissignal dürfen die Stimmlippen dabei nicht zu stark aneinander kleben [2]. Im Silikon können zusätzlich magnetisch polarisierbare Partikel eingegossen werden, sodass dieses ein magnetorheologisches Elastomer bildet [1]. Über ein induziertes Magnetfeld kann so die Festigkeit des Silikons gesteuert werden, wodurch die Grundtonhöhe moduliert wird.

Fazit und Ausblick

Mit dem Projekt *Faszination Sprechende Maschinen* konnten die Exponate der HAPS in ihrem historisch-kulturellen Kontext neu bewertet und eingeordnet werden, sodass sie nun in Lehrveranstaltungen und Praktika als Demonstrationsobjekte einsetzbar sind. Gleichzeitig wird die Sichtbarkeit der von der Kustodie betreuten Sammlungsobjekte der Technischen Universitä Dresden durch ihre multimediale Präsentation deutlich erhöht. Darüber hinaus können Entwicklungen im Bereich der Sprachsynthese-Technologien mit ihren Traditionsbezügen nun auch wahrnehmbarer in die Zukunft projiziert werden. Zukünftig soll die Erfassung der HAPS fortgeführt werden und der Quellcode der Virtuellen Ausstellung soll offengelegt werden. Der Sprechdemonstrator *mika*[2] wird mit einer ansteuerbaren künstlichen Zunge erweitert. Zudem werden einige HAPS-Objekte in der Sonderausstellung *DER SCHLÜSSEL ZUM LEBEN – 500 Jahre mechanische Figurenautomaten* (25. Juli 2020 – 10. Januar 2021, Japanisches Palais in Dresden) von den Staatlichen Kunstsammlungen Dresden präsentiert.

Abbildung 3: Sprechapparat mika^2 als Forschungsplattform.

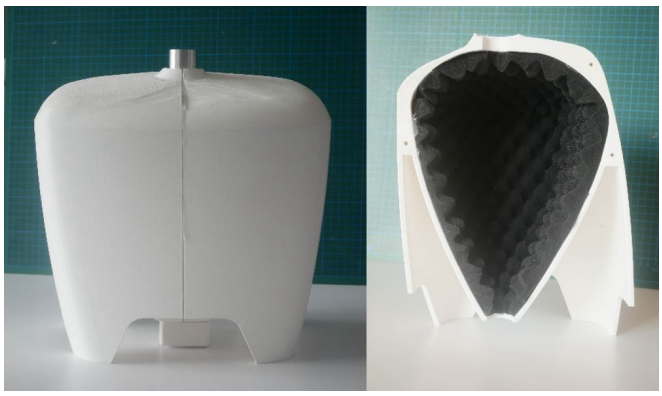

Abbildung 4: 3D-Druck des künstlichen Lungenvolumens – von innen gedämmt *(rechts)*.

Danksagung

Ein herzlicher Dank gilt Kirsten Vincenz und Jörg Zaun von der Kustodie der Technischen Universität Dresden für die enge und intensive Zusammenarbeit in diesem Projekt.

Das Verbundprojekt *Faszination Sprechende Maschine: Technologischer Wandel der Sprachsynthese über zwei Jahrhunderte* (01UQ1601 A|B) wurde vom Bundesministerium für Bildung und Forschung gefördert.

Literatur

[1] Dohmen, E.; Borin, D.; Gabriel, F.; Odenbach, D.; Birkholz, P.: Artificial vocalis muscles for speech synthesis made from contact-less adaptable magnetorheological elastomer. In: Proc. of the International Electorheological Fluids and Magnetorheological Suspensions Conference (ERMR2018), Maryland, USA.

[2] Dohmen, E.; Borin, D.; Birkholz, P.; Gabriel, F.; Häsner, P.: Surface stickiness and waviness of two-layer silicone structures for synthetic vocal folds. In: Birkholz, P.; Stone S. (Hrsg.): Elektronische Sprachsignalverarbeitung 2019. Dresden: TUDpress 2019 (Studientexte zur Sprachkommunikation, Bd. 93), S. 221 – 230.

[3] Engeln, L.; Groh, R.: VirtEx: Eine Ontologie-basierte Virtuelle Ausstellung. Mensch und Computer 2018, Tagungsband.

[4] Engeln, L.; Hube, N.; Groh, N.: Immersive VisualAudioDesign: Spectral Editing in VR. In: Proc. of the Audio Mostly 2018 on Sound in Immersion and Emotion (AM'18). ACM, New York, NY, USA, Article 38, 4 pages, https://doi.org/10.1145/3243274.3243279

[5] Hoffmann, R.; Jäckel, R.: Zur Geschichte des Vocoders in der Sowjetunion. In: 44. Jahrestagung für Akustik, DAGA 2018, München, Tagungsband S. 840 – 843.

[6] Hoffmann, R.; Birkholz, P.; Gabriel, F.; Jäckel, R.: From Kratzenstein to the Soviet vocoder: Some results of a historic research project in speech technology. In: Karpov, A.; Jokisch, O.; Potapova, R. (Hrsg.): Speech and Computer. 20th International Conference SPECOM, Leipzig 2018, Proceedings. Springer 2018 (Lecture Notes in Artificial Intelligence; 11096), S. 215 – 225.

[7] Hoffmann, R.: 50 Years Institute of Acoustics and Speech Communication, 30 Years Conference Electronic Speech Signal Processing, 20 Years Historic Acoustic-phonetic Collection. In: Birkholz, P.; Stone S. (Hrsg.): Elektronische Sprachsignalverarbeitung 2019. Dresden: TUDpress 2019 (Studientexte zur Sprachkommunikation, Bd. 93), S. 1 – 8.

[8] Hoffmann, R.: Raum- und Schiff-Phonetik – Panconcelli-Calzia und der Lautstärkemesser nach Barkhausen. In: 45. Jahrestagung für Akustik, DAGA 2019, Rostock, Tagungsband S. 690 – 693.

[9] Hofmann, E.: Das „Sprechende Bilderbuch“ und sein Erfinder Theodor Brand. Dresden: TUDpress 2018 (Studientexte zur Sprachkommunikation Bd. 92).

[10] Mehnert, D.: Historische phonetische Geräte: Katalog der historischen akustisch-phonetischen Sammlung (HAPS) der Technischen Universität Dresden, Teil 1. Dresden:TUDpress 2012 (Studientexte zur Sprachkommunikation, Bd. 62).