EINE COMPUTER–LERNHILFE FÜR DAS LIPPENLESEN

Dipl.–Ing. Hans–H. Bothe
Prof. Dr.–Ing. Dietrich Naunin
Institut für Elektronik, TU Berlin

1. Problematik

Menschen mit defektem auditiven System nehmen einen großen Teil der sprachlichen Information durch das Absehen der bewegten Mundregion des Sprechers auf. Durch schnelles und folgerichtiges Kombinieren müssen die koordinierten Mundbewegungen mit dem vorhandenen Inventar von sprachlichen Zeichen in Zusammenhang gebracht werden. Besonders bei frisch Ertaubten muß die Verarbeitung der Eingangsinformation über die beteiligten Kommunikatonskanäle (Augen, Ohren) vollständig umstrukturiert werden. In Anbetracht von über 200.000 Spätertaubten und Gehörlosen sowie über einer Million Hörgeräteträgern allein in der Bundesrepublik stehen jedoch für den (besonders in der Anfangsphase des Lernens) notwendigen Einzelunterricht entschieden zu wenig Pädagogen zur Verfügung.

Im Rahmen eines Forschungsprojekts des Instituts für Elektronik der TU Berlin wurde deshalb das Rahmenkonzept einer Lern– und Trainingshilfe für den Absehunterricht entwickelt, das die Arbeit des Pädagogen möglichst effektiv ergänzen und schnell zu einem gesicherten Erkennen der charakteristischen Bewegungsabläufe führen kann.

Dazu kann ein beliebiger Text in phonetischer Form mit Hilfe der Tastatur oder einer Maus in den Computer eingegeben und auf dem Bildschirm in den Trickfilm eines sich bewegenden Gerichts umgesetzt werden. Die wichtigen Bestandteile werden in ihrer Bewegung deutlich hervorgehoben. Neben variabler Ablaufgeschwindigkeit und beliebig häufiger Widerholbarkeit können zusätzlich Sprachmerkmale wie Stimmhaftigkeit und Betonung, die sich im optischen Erscheinungsbild nicht oder nur unvollkommen äußern, in mehreren Hilfsstufen eingeblendet bzw. taktil mitgeteilt werden. Optional kann das System auch mit einer Sprachausgabeeinheit versehen werden.

Die Lernhilfe ist als technisches Hilfsmittel für Gehörlosenpädagogen und Logopäden konzipiert und kann daneben auch als individuell nutzbares Sprachlabor für das Absehen und zum autodidaktischen Lernen eingesetzt werden. Damit eröffnen sich für Hörgeschädigte die folgenden Perspektiven:

- Verbesserung der individuellen Trainingsmöglichkeiten bei freier Zeiteinteilung unter Anleitung eines Lehrers, z.B. Weiterüben von individuellen Problemstellungen nach Unterrichtsschluß,
- mehr Sicherheit im beruflichen Alltag durch gezieltes Trainieren von berufs– und situationsspezifischen Schlüsselwörtern.

Eine ergänzende Anwendung findet sich beim Einsatz im Deutschunterricht für Ausländer.

2. Segmentierung der Mundbewegungen

Die Bewegungen der sichtbaren Teile des Artikulationstrakts lassen sich als Ausdruck einer visuellen Sprache begreifen. Seit der bahnbrechenden Arbeit von Alich [1] ist davon auszugehen, daß sich diese über die Annahme einer bestimmten Anzahl von segmental minimalen Einheiten analytisch erfassen läßt. Die sprecherunabhängigen Gestaltklassen werden in Analogie zur Segmentierung der Sprache als "Kineme" bezeichnet. Im Deutschen unterscheidet man acht konsonantische und vier vokalische Kineme, die im folgenden aufgeführt sind.

Konsonantische Kineme:

1.	Bilabiales Kinem (B)	/p, b, m/
2.	Stationäres Bilabiales Kinem (M)	/m/ (in Finalposition)
3.	Labiodentales Kinem (F)	/f, v/
4.	Dentales Kinem (D)	/s, z, t, d, n/
5.	Koronales Kinem (L)	/t, d, n, l, r/
6.	Dorsales Kinem (C)	/c, j/
7.	Gutturales Kinem (G)	/x, k, g, ng, h/
8.	Gerundetes Dentalkinem (S)	/sch/

Vokalische Kineme:

1.	weites Palatalkinem (A)	/a, a:, e:, ɛ, ɛ:/
2.	enges Palatalkinem (I)	/e, i, i:, ɛ, ɛ:, ə/
3.	weites Velarkinem (O)	/o, o:, ø, œ, ɔ/
4.	enges Velarkinem (U)	/u, u:, y, y:/

Beim Sprechen von Lautfolgen führen die beteiligten Sprechorgange bestimmte, koordinierte Bewegungen aus, die durch eine Überlagerung von aktiv geformten Ausprägungen und passiven Mitwegungen entstehen. Vorbereitende Analysen dieser Vorgänge wurden beispielsweise von Menzerath [2] und Lindner [3] durchgeführt. Ausgehend von der Aussprachenorm der deutschen Standardaussprache hat Lindner das akustische Material in lautliche Zweiersequenzen aufgespalten und die Einstellungen der Sprechorgane als für die Aussprache des entsprechenden Lautes wesentlich, wahrscheinlich oder frei unterschieden. Die Uneindeutigkeiten entstehen durch koartikulatorische Verflechtungen, bei denen bestimmte Artikulationsparameter jeweils noch nacheilen bzw. schon vorbereitend wirken. Für eine realistische Computeranimation sind die beschriebenen Effekte von entscheidender Bedeutung. Eine quantitative Bestimmung kann mit Hilfe von Verwechslungsmatrizen zur Abbildung von Phonemfolgen auf Kinemfolgen bei Modulation der Lautdauer durchgeführt werden.

3. Entwicklungskonzept

Die Ausprägung der Artikulation wird von verschiedenen, übergeordneten Faktoren wie Akzentuierung, Dialekt und persönlicher Sprachrhythmik beeinflußt. Das Problem einer umfassenden Bewegungsanalyse würde damit sehr komplex, wenn nicht unrealistisch. Daher sind zunächst prototypische Sprecher mit deutscher Standardaussprache heranzuziehen. Bei Reduktion des Ausgangsmaterials auf spezifische Parameter geschieht die Erarbeitung von wirklichkeitsnahen Modellen mit Hilfe von Videofilmen. Schon bei der Analyse des empirischen Materials werden die digitalisierten Videobilder in jeweils ein Hintergrundbild und eine a priori definierte Menge von beweglichen Elementen innerhalb dieses Bezugsbildes zerlegt (Lippen, Zungenspitze, Unterkiefer mit Schneidezähnen samt Kinnpartie). Die beweglichen Elemente lassen sich mit Hilfe von Umrißkonturen

im Hintergrundbild lokalisieren und bei der späteren Animation in Ahängigkeit von weiteren Parametern (z.B. Vorwölbung der Lippen) mit Hilfe von mathematischen Beleuchtungsmodellen schattieren.

4. Computeranimation

4.1 Prinzipielles

Die momentan verfügbare Version der Lernhilfe wurde auf einem Kleinrechner vom Typ Atari ST implementiert. Sie stellt ein Rahmenkonzept dar, in das die gewonnenen Erkenntnisse über Koartikulationseffekte zu integrieren sind. Eine Implementation von Regeln für Diphone und Triphone mit Hilfe von Verwechslungsmatrizen ist vorbereitet. Ferner ist geplant, in einem weiterführenden Forschungsprojekt Übergangsregeln für syllabische Strukturen zu erarbeiten und in das Animationsmodell zu integrieren.

Das Programm basiert zunächst auf phonembezogenen Mundbildern unter Einbezug von Lippen, Zunge(nspitze) und Zähnen. Im Arbeitsspeicher des Rechners liegt ein statisches Gesichtsbild, der Bewegungsablauf entsteht durch aufeinanderfolgendes Einblenden von Bildausschnitten. Diese werden nach Anwahl einer neuen Phonemfolge zunächst berechnet. Neben der Mundregion erscheinen auch Augen, Augenbrauen, Nasenflügel, und Kehlkopfregion beweglich, um Satzende, Betonung, Nasalität und Stimmhaftigkeit anzuzeigen und den ansonsten starren Gesichtsausdruck aufzulockern.

Koartikulationseffekte werden dabei momentan (nur) insofern berücksichtigt, als die vorgegebenen Mundbilder bei der Trickfilmsynthese in der Mitte der Lautübergänge angeordnet sind, welche damit von den Nachbarphonemen beeinflußt werden. Die Verwechslungsmatrizen stellen eine 1:1–Transformation zwischen Phonemfolgen und Mundbildern dar, so daß systembedingte Fehler auftreten. Diese können jedoch selbständig mit Hilfe eines integrierten, komfortablen Graphikeditors durch Variation der Konturen und der zeitlichen Übergangsfunktionen zwischen den Lauten korrigiert werden. Dabei beschäftigt sich der Benutzer in einem interaktiven Prozeß mit den Einflußfaktoren für "richtig" koordinierte Bewegungsabläufe und trainiert zusätzlich Perzeptionsfähigkeit und Aufmerksamkeit.

4.2 Benutzeroberfläche

Durch eine konsequent graphische Gestaltung der Benutzeroberfläche können auch im Umgang mit Computern unerfahrene Anwender die Bedienung des Programmes rasch erlernen. Alle ausführbaren Funktionen sind entweder durch Textboxen oder durch symbolische Bilder mit der Maus anwählbar. Das Programm gliedert sich in vier Teile,

das SPRECH–Modul zum Vorsprechen von Phonemfolgen,
das EINGABE–Modul zum Eingeben von Phonemfolgen,
das MUND–Modul zum Ändern von Mundkonturen und Übergangsfunktionen,
das EXTRAS–Modul zum Ändern des graphischen Erscheinungsbildes.

Von jedem Programmodul sind jeweils die anderen drei aufrufbar.

4.3 Generieren der Zwischenbilder

Die zu berechnenden Bildausschnitte der Mundregion haben eine Größe von 128*128 Pixeln und setzen sich aus der Zunge, den Zähnen, den Lippen und einem Teilausschnitt des statischen Hintergrundbildes zusammen. Während die Lippenkonturen in Form von Polygonkoordinaten im Speicher vorliegen, existieren für Zunge und Zähne vorgefertigte Bilder. Das Bild der Zunge ist 256 Pixel hoch, in Abhängigkeit von der jeweiligen Stel-

lung der Zungenspitze wird der entsprechende Bildausschnitt verwendet. Die Zwischenbilder werden mit Hilfe von Interpolationsverfahren für die Stützpunkte und die Offsetwerte von Zungen– und Zahnstellung berechnet.

Es werden binäre Schwarzweißbilder dargestellt, wobei Graustufen durch verschiedene Punktdichten entstehen. Damit wird eine Trickfilmgenerierung in akzeptabler Zeit möglich, weil mit einer einzigen Assemblerroutine gleichzeitig 32 Pixel verändert bzw. bewegt werden können.

4.4 Zeitliche Übergangsfunktionen

Ein Lautübergang besteht aus acht Einzelbildern. Bei normaler Sprechgeschwindigkeit von fünf bis zehn Lauten pro Sekunde entsteht bei der Animation eine Bildfolgefrequenz von 40 bis 80 Bildern pro Sekunde. Die eigentlichen Stützbilder liegen dabei eingebettet in die Zwischenbilder, so daß der (logische) Lautübergang eine Ein– und Ausklingphase erhält und für jedes Phonem typische Bewegungsprofile mit Hilfe von graphischen Schiebereglern eingestellt werden können (Typendarstellung von Vokalen, Explosivlauten,...). Koartikulationseffekte bezüglich der Lautdauer können durch Modulation der Vorgabewerte berücksichtigt werden.

4.5 Zungenbewegung

Die Bewegung der Zungenspitze erfolgt zum Teil unabhängig von der Bewegung der Lippen. Anfangs– bzw. Zielstellung werden durch jeweilige Zungenoffsets repräsentiert, deren zeitliche Zuordnung innerhalb eines Lautübergangs für jedes Phonem eingestellt wird (Zunge bewegt sich nur innerhalb eines Bereiches von Bild i bis Bild j). Die Vorgabewerte sind in Abhängigkeit vom Kontext dynamisch veränderbar.

4.6 Speicherverwaltung

Der zur Verfügung stehende Arbeitsspeicher des Rechners wird in Bildblöcke für die einzelnen Lautübergänge eingeteilt, und zwar an dieser Stelle von Stützbild zu Stützbild. Damit können bereits berechnete Lautübergänge zum Darstellen einer nächsten Phonemfolge mit einer relativ großen Wahrscheinlichkeit wiederverwendet werden, und die Berechnungszeiten für den Gesamtfilm verringern sich. Vorhandene Bilddaten sind sowohl für Übergänge vom Quell– zum Zielphonem als auch umgekehrt verwendbar.

5. Weiterführende Perspektiven

Wie schon oben erwähnt, sollen Untersuchungen über das Zusammenspiel von umfassenderen Bewegungseinheiten angestellt werden, um die Einflußzonen der Artikulationsparameter auf einen "natürlichen" Rahmen zu erweitern. Als Minimalsegmente bieten sich dabei zunächst syllabische Strukturen an, die als grundlegende Einheiten der Sprachproduktion anzusehen sind.

Literaturangaben:

[1] Alich, Georg: Zur Erkennbarkeit von Sprachgestalten beim Ablesen vom Munde (Dissertation) Bonn 1960/61.

[2] Menzerath, Paul und A. de Lacerda: Koartikulation, Steuerung und Laut–abgrenzung. Berlin und Bonn 1933.

[3] Lindner, Gerhart: Entwicklung von Sprechfertigkeiten bei gehörlosen Kindern. Berlin: Volk und Wissen 1984.