Institut für Informatik

Informatik in Bildung
und Gesellschaft


# Die Geschichte der Sprachsynthese anhand einiger ausgewählter Beispiele

Arne Hoxbergen

6. Juni 2005

Betreuer: Dr. Jochen Koubek

Seit langen schon versuchen die Menschen ein perfektes Abbild ihrer selbst zu erzeugen. Schon aus der Antike sind uns sprechende Köpfe bekannt, mit denen Priester die Gläubigen beeindruckten. Während langer Zeit war durch das Diktat der christlichen Kirche jede Bemühung, in dieser Richtung zu forschen und zu arbeiten, unterdrückt worden. So soll Albertus Magnus einen sprechenden Kopf gebaut haben, der dann aber von einem seiner Schüler als herätisches Werk zerstört wurde. Erst mit dem Barrock und der Befreiung der Wissenschaft von den Zwängen der Religion erstarken die Bemühungen zur Schaffung künstlicher Abbilder. Dies mag nicht unwesentlich auf die damalige Weltsicht zurückzuführen sein, nach der die Welt als Ganzes aus kleinen, atomaren Bausteinen besteht, die jedes für sich mechanisch vollständig beschreibbar sind. Folgerichtig war auch die Welt als Ganzes und alle ihre Teile, also auch der Mensch, als das Zusammenwirken einzelner mechanischer Teile beschreibbar. Das Uhrwerk wurde zum Symbol dieser Anschauung. Vom Planetensystem bis zum Menschen, versucht man alles als Maschine zu sehen. Es ist der Erbauer des schachspielenden Türken, Wolfgang von Kempelen, der dann auch die erste Sprechende Maschine baut, die uns überliefert ist. Seine Arbeiten zur Sprache sollen über die nächsten Jahrhunderte das Grundgerüst der Sprechmaschinen darstellen.

Erst mit der Elektrotechnik kommen andere neue Konzepte zur synthetischen Sprache auf, die ihrerseits die Sprachforschung beeinflussen. Versucht man anfänglich die mechanischen Maschinen elektronisch nachzugestalten, so wird man doch bald eigene Modelle der Sprache aufstellen, die ihrerseits neue Erkenntnisse in der Sprachforschung bedingen. Schon seit Kempelen gibt es diese enge Verbindung. Doch die Elektrotechnik bringt eine neue Geschwindigkeit in die Forschung. Es ist die Analogie zwischen der Schallwelle und der elektrischen Welle, die dies auslöste, und bis heute anhält.

Mit der Computertechnik gab es dann eine erneute Beschleunigung. Kann der Rechenknecht doch dank seiner hohen Geschwindigkeit, seines enormen Speichervermögens und seines schnellen Rekombinierens von Informationen Sprachsegmente dynamisch zusammensetzten. Die elektronischen analogen Schwingungen werden zu Bitmustern, die beliebig kombinierbar sind. Und es ist möglich, Sprache in Echtzeit zu erzeugen. Der Ansatzpunkt verschiebt sich nun von der Erzeugung sprachlicher Laute, und Lautfolgen in einer vorgegebenen Reihenfolge hin zur Echtzeit-Erzeugung der Reihenfolge selbst. Diese Studienarbeit soll diese Entwicklung der Technik und der zugrundeliegenden Erkenntnisse aufzeigen, und es so dem interessierten Leser ermöglichen, einige Eigenarten dieser Techniken und der daraus enstehenden Anwendungen zu verstehen.

# Inhaltsverzeichnis

# 1 Eine kleine Phonetik

Ich möchte kurz einige grundsätzliche Begriffe klären, die in der folgenden Arbeit häufiger auftauchen werden. Dies sind insbesondere Begriffe, die in der Sprachforschung vorkommen.

## 1.1 Phonetik

> *Die Phonetik (auch Lautlehre) ist die Lehre der von Menschen hervorgebrachten Laute. Man unterscheidet:*
>
> ** artikulatorische Phonetik ("Sprechakt-Lautlehre"), die untersucht, welche und wie Laute beim Sprechen erzeugt werden,*
>
> ** akustische Phonetik, die die physikalische Natur der Schallwellen untersucht, die die Laute bilden, und*
>
> ** auditive Phonetik, deren Forschungsgebiet die Vorgänge beim Empfang der Laute im menschlichen Ohr sind.*[1]

In der Sprachforschung versucht man wie in anderen Forschungsgebieten auch, die kleinsten Teilchen zu finden. Derer gibt es zwei Arten, die voneinander unterschieden werden müssen, Phone und Phoneme. Beide spielen in der Sprachsynthese eine Rolle.

[1] de.wikipedia.org, *Phonetik*.

## 1.2 Phon

> *In der Linguistik ist ein Phon (seltener auch Fon geschrieben) das kleinste Element einer sprachlichen Äußerung.[. . . ]Phone sind lautliche Repräsentationen von Phonemen, mit denen sie nicht verwechselt werden dürfen.*[2]

Phone sind also kleinste Lauteinheiten, solche die nicht in kleinere, eigenständig erzeugbare Laute zerlegt werden können. Diese müssen sich akustisch voneinander unterscheiden lassen. Phone können auch Phoneme sein.

## 1.3 Phonem

> *Ein Phonem (auch Fonem geschrieben) ist die kleinste bedeutungsunterscheidende Einheit eines Sprachsystems (gemäß dem strukturalistischen Modell einer Sprache) und der wissenschaftliche Untersuchungsgegenstand der Phonologie.*
>
> *Es handelt sich hierbei um die von den Sprechern einer Sprache als bedeutungsunterscheidend identifizierten Einzellaute.*[3]

Die Betonung liegt hierbei auf der Bedeutungsunterscheidung. In der deutschen Sprache ist das */ch/* mal wie in *ich* oder mal wie in *ach* gesprochen. Wenn man das Wort *ich* mit einem *ach-/ch/* spricht, ändert sich nicht die Bedeutung des Wortes. Die beiden *ch* sind also *„phonetisch realisierte Varianten eines einzigen Phonems. Sie sind allerdings, bevor man sie als Varianten eines Phonems klassifizieren kann, zunächst zwei verschiedene Phone: [ç] und [χ].”*[3] Solche Laute nennt man auch Allophone.

[2] de.wikipedia.org, *Phon*.
[3] de.wikipedia.org, *Phonem*.

## 1.4 Vokal

> *Ein Selbstlaut oder Vokal [...] ist ein Laut, bei dessen Artikulation der Phonationsstrom weitgehend ungehindert ausströmen kann. Vokale sind stimmhaft.*[4]

Vokale lassen sich daher sehr gut durch Resonanzkörper, ähnlich Orgelpfeifen nachahmen. Im deutschen gibt es, je nach Zählweise, 15 bzw. 16 Vokale, wobei jeweils eine lange und eine kurze Version durch den gleichen Buchstaben geschrieben werden. Fünf Vokale und drei Umlaute, wobei das kurze */ä/* und das kurze */e/* ein Phonem sind (z. B. Lärche und Lerche).

## 1.5 Konsonant

> *Unter Konsonant [...] versteht man einerseits einen Laut, dessen Artikulation eine Verengung des Stimmtraktes beinhaltet, so dass der Atemluftstrom ganz oder teilweise blockiert wird und es zu hörbaren Turbulenzen (Luftwirbelungen) kommt. Konsonanten sind Hemmnis überwindende Laute.*[5]

Konsonanten werden in stimmhafte und stimmlose unterschieden. Ein stimmhafter ist das */m/*, ein stimmloser das */f/*.

## 1.6 Frikative

> *Ein Frikativ (auch Reibelaut, Engelaut, Konstriktiv, Spirans, Spirant) ist ein nach seiner Artikulationsart benannter Konsonant.*[6]

Ein Reibelaut wird erzeugt, indem der Luftstrom eingeengt wird, so daß an der Engstelle die Luft verwirbelt wird, und durch diese Verwirbelung der Laut ensteht.

[4] de.wikipedia.org, *Vokal*.
[5] de.wikipedia.org, *Konsonant*.
[6] de.wikipedia.org, *Frikativ*.

## 1.7 Diphthong

> *Ein Diphthong* [...] *ist ein Doppellaut/Zwielaut aus zwei Vokalen. Ein Vokal geht dabei in den anderen über*[...][7]

Unter die Diphthonge fallen Zwischenlaute und Umlaute wie /aɪ/ in Weise, /a$\underset{\frown}{\text{ʊ}}$/ in Haus und /ɔ$\underset{\frown}{\text{Y}}$/ in Heu.

## 1.8 Diphon

> *Ein Diphon beschreibt in der konkatenativen Sprachsynthese den kurzen Abschnitt (Baustein) gesprochener Sprache, der in der Mitte eines Phons beginnt und in der Mitte des folgenden Phons endet.*[8]

Diphone können also als Lautübergänge angesehen werden.

## 1.9 Vokaltrakt

> *Der Vokaltrakt umfaßt den gesamten Rachen- und Mundraum von den Stimmlippen an aufwärts. Seine Hauptaufgabe ist der Transport von Nahrung und Atemluft. Die sekundäre Funktion ist die der Resonanz und Artikulation. Anatomie und Funktion* [sic]*Der im Kehlkopf entstandene Primärklang erfährt im Vokaltrakt eine Veränderung durch Verstärkung oder Abdämpfung einzelner Frequenzen. Größe und Form des Vokaltraktes sind sehr veränderlich: Die Kehlkopfstellung bestimmt seine Länge, die Beweglichkeit von Mundraum (Zunge, Kiefer, Lippen) und Rachenraum beeinflußt seine Form.*[9]

[7] de.wikipedia.org, *Diphthong*.
[8] de.wikipedia.de, *Diphon*.
[9] Ciba, *Funktionales Stimmtraining 1*.

## 1.10 Formant

> *Formanten (von lateinisch formare = formen) nennt man bei Musikinstrumenten oder der menschlichen Stimme Frequenzbereiche, bei denen die Lautstärke angehoben ist. (Maxima im Spektrum).*[10]

Formanten ähneln den sogenannten spektralen Kennlinien, wie sie bei elektrischen Bauteilen vorhanden sind. Diese Analogie wird in einigen elektrischen Sprechmaschinen benutzt.

[10] de.wikipedia.org, *Formant*.

# 2 Sprechmaschinen

## 2.1 Die mechanischen Sprechmaschinen

Bereits in der Antike gab es Versuche, sprachfähige Abbilder des Menschen zu erschaffen. Dies wurde auf äußerst einfache Weise realisiert. Man legte einfach ein Rohr von einem Raum unterhalb einer Statue durch das Innere zum Mund der Statue. Diese Sprachrohre wurden noch lange Zeit verwendet, z.B. in der Schiffahrt. Dieser Ansatz ist allerdings keine Sprachsynthese.

Während des Mittelalters wurde durch das Diktat der Kirche die Wissenschaft so sehr eingeengt, gar eingespannt, daß eine freie forschende Tätigkeit praktisch nicht mehr möglich war:

> *The earliest speaking machines were perceived as the heretical works of magicians and thus as attempts to defy god. In the thirteenth century the philosopher Albertus Magnus is said to have created a head that could talk, only to see if destroyed by St. Thomas Aquinas, a former student of his, as an abomination. The English scientist-monk Roger Bacon seems to have produced one as well. That fakes were appearing in Europe in the late sixteenth and early seventeenth centuries is shown by Miguel de Cervantes's description of a head that spoke to Don Quixote – with the help of a tube that led to the floor below. Like Magnus, this fictitious inventor also feared the judgement of religious authorities, though in his case he took it upon himslef to destroy the heresy. By the eighteenth century science had started to shed its connection to magic, and the problem of artificial speech was taken up by inventors of a more mechanical bent.*[11]

[11] Lindsay, *Talking Head*; zitiert nach: Rubin / Vatikiotis-Bateson, *Kratzenstein*.

Mit dem Barrock und der Aufklärung ändert sich die Lage. Am Anfang stehen hierbei die Herren Kratzenstein und von Kempelen, die unabhängig voneinander erste Ansätze zur Sprachsynthese umsetzten. Obwohl von Kempelen schon vor Kratzenstein mit seiner Arbeit beginnt, wurde seine Maschine erst viele Jahre später fertig.

### 2.1.1 Kratzensteins Resonatoren

1779 nahm Christian Gottlieb Kratzenstein an der Ausschreibung zum Jahrespreis an der kaiserlichen Akademie von St. Petersburg teil, die er gewann. Er modellierte und baute eine reihe akkustischer Resonatoren, die die fünf Vokale */a/*, */e/*, */i/*, */o/*, */u/* nachahmten. Die Universität hatte den Preis für die Erklärung der physischen Unterschiede bei der Erzeugung der Laute ausgeschrieben. Kratzenstein baute seine Resonatoren, indem er den menschlichen Vokaltrakt nachbildete. Durch Variation der Resonatoren konnte eine begrenzte Nachahmung von Worten erfolgen.[12]

### 2.1.2 Von Kempelens sprechende Maschine

Unabhängig von Kratzenstein entwickelte Wolfgang von Kempelen eine Maschine zur Nachahmung von Sprache. Er erforschte die menschliche Sprache und versuchte seine Erkenntnisse anhand dieser Maschine zu bestätigen. Er begann mit der Arbeit an seiner Maschine bereits 1769. Obwohl er mit dieser Maschine seiner Zeit weit voraus war, blieb sie weitestgehend unbeachtet. Berühmt wurde von Kempelen, Hofingenieur am Hofe Maria Theresias, für eine Spielerei, eine Illusion. Er baute einen äußerst komplizierten Pseudoautomaten in Form eines Türken. Dieser Automat sollte Schach spielen können. Er führte diesen Automaten öffentlich vor, so wie ein Zauberkünstler seine Stücke präsentiert. Zu Beginn der Vorstellung öffnete er den Automaten und ließ das Publikum einen Blick hineinwerfen. Durch die Verwendung von Spiegeln nahmen die Zuschauer an, daß im Inneren des Automaten kein Platz für einen Menschen sei. Obwohl

[12] Rubin / Vatikiotis-Bateson, *Kratzenstein*.

von Kempelen am Beginn seiner Vorführung erklärte, es handle sich um eine Täuschung, erlagen viele Menschen der Illusion, es sei eine Maschine geschaffen worden, die dem menschlichen Geiste ebenbürtig sei. Mit zunehmender Popularität geriet von Kempelen zunehmend unter Erklärungsnot. 84 Jahre lang erstaunte und polarisierte der Automat die Massen, bis er 1854 in Philladelphia beim Brand im Chineese Museum zerstört wird.[13]

Während der „getürkte" Schachautomat, der tatsächlich von einem kleinen Menschen gesteuert wurde, das allgemeine Bild von Kempelen als *„Randfigur der Geschichte"*(Theodor Heuss, zitiert nach[13]) prägt, sind seine Arbeiten im Bereich der Sprachforschung jedoch bis in die heutige Zeit hinein bedeutend.

Wolfgang von Kempelen beschrieb seine Erkenntnisse zur menschlichen Sprache zusammen mit seiner sprechenden Maschine in seinem Buch, das 1791 unter dem Titel *„Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine"* erschien. Es sollten andere Menschen diese Maschine nachbauen und verbessern können.

> *Seine sprechende Maschine kann auch als eines der ersten Open-Source Projekte angesehen werden, da er sie detalliert beschrieb, um Interessierten einen Nachbau zu ermöglichen und vor allem die Weiterentwicklung zu fördern.*[14]

Kempelen mußte zur Erschaffung einer Maschine, die nicht nur einzelne Töne, sondern ganze Lautfolgen, gar Wörter und Sätze erzeugen konnte, erst die theoretischen Grundlagen der Sprache erforschen. Der Bau und die ständige Weiterentwicklung der Maschine und die Erforschung der Sprache bedingten sich dabei gegenseitig. Er wird daher auch als einer der ersten *„experimentellen Phonetiker"* bezeichnet.[15] Diese Erforschung der Sprache machte einen sehr großen Anteil der gesamten Zeit aus, die er für den Bau der Maschine aufwendete:

---

[13] Strouhal, *Kempelens Türke: Eine Schach-Methaphern-Maschine aus dem Spätbarrok*.
[14] lexikon.eventax.de, *Wolfgang von Kempelen*.
[15] Traunmüller, *Wolfgang von Kempelens sprechende Maschine*.

> *He worked on his machine for more than 14 years that where preceded by by a long time spent doing research on the articulation physiology and accoustics of speech sound.*[16]

Die sprechende Maschine besteht aus drei wesentlichen Bestandteilen. Einem Blasebalg, der die Funktion der Lungen übernimmt. Dieser wird mit dem rechten Unterarm bedient, wobei das „Einatmen" durch ein Gegengewicht ausgeführt wird. Des weiteren besteht sie aus einem Tonerzeuger. Hier wird der Luftstrom über ein Rohrblatt geführt, dessen Schwingungen den Grundton erzeugen. Dieses Rohrblatt war aus Elfenbein gefertigt. Zusätzlich gab es einen zweiten Luftweg, der durch ein parallel geführtes Rohr zum Mund führte. Damit war es möglich den Luftdruck im Mund der Maschine ansteigen zu lassen, wenn man stimmlose Konsonanten erzeugen wollte. Der dritte Teil war die Nachbildung des Vokaltraktes. Der Mund der Maschine war ein trichterförmiges Rohrstück. Dieses Ansatzrohr ist das Haupt-Steuerelement. Durch geeignetes teilweises Verdecken kann die Resonanzeigenschaft so verändert werden, daß Vokale erzeugt werden. Zur Bildung von stimmlosen Konsonanten muß das Ansatzrohr ganz verdeckt werden. Ein zusätzlicher kleiner Blasebalg bringt dann ein Verpuffungsgeräusch hervor. Die Nase des Gerätes kann verwendet werden, um nasale Laute zu erzeugen. Bei nichtnasalen Lauten wird sie verschlossen.

Mit der rechten Hand kann man zwei kleine Hebel bedienen, die Frikative bilden. Dazu bediente sich Kempelen zweier kleiner Pfeifen, die Zischlaute erzeugten. Somit können die Laute /s/ und /[ʃ]/ sowie /z/ und /[ʒ]/ hervorgebracht werden. Ein dem /r/ ähnliches Geräusch kann erzeugt werden, indem ein Stück Draht auf das Rohrblatt gelegt wird. Dadurch entsteht ein rasselnder Ton. Es gibt auch Töne, die nicht dargestellt werden können. Dies sind /d/, /t/, /k/, und /g/. Die Maschine hat kein Analogon zur Zunge. Ein /l/ läßt sich imitieren, indem man einen Finger in den Mund steckt.[15]

Obwohl oder gerade weil Kempelens Maschine nicht alle Töne der in Europa verwendeten Sprachen nachahmen konnte, war sie so wichtig für die Geschichte der Sprachsynthese. Dieser Mangel an Funktion provozierte

[16] Gósy, *On the Early History of Hungarian Speech Research*,S. 156.
[15] Traunmüller, *Wolfgang von Kempelens sprechende Maschine*.

Kempelen zu weiteren Nachforschungen. So stellte er schließlich fest, daß es gar nicht nötig war, alle Laute exakt wiederzugeben, da unser Gehör sich leicht täuschen lässt. Diese Erfahrung wird erst viel später durch erste empirische Nachforschungen im Bereich der Wahrnehmungsforschung methodisch abgesichert.

Kempelen erkannte, daß sich unterschiedliche Sprachen unterschiedlich gut synthetisieren lassen. Obwohl er dies auf seine Maschine zurückführte, gilt dies bis heute für alle Verfahren, wobei natürlich für unterschiedliche Verfahren diese Reihenfolgen sich ändern können. Kempelen schreibt zu seiner Maschine: „[...]*besonders wenn man sich auf die lateinische, französische oder italienische Sprache verlegt, denn die deutsche ist* [...] *um sehr vieles schwerer.*"[17]

Kempelens Maschine stellte für die nächsten einhundert Jahre den Stand der Technik dar, bis auf einige Nachbauten mit geringfügigen Veränderungen tat sich fast nichts. Es handelt sich um eine mechanische Nachbildung des Sprechapparates. Kempelen konnte sich, da er die gleichen Prinzipien wie die Natur beim Original benutzt, auf die Vereinfachung (im Sinne einer Abstraktion) beschränken. Dies machte es ihm möglich, relativ große Fortschritte in kurzer Zeit zu erzielen. Die genaue Entstehungsgeschichte, einschließlich der auftretenden Probleme, wird in seinem Buch dargetsellt.

Es ist ein Exemplar der letzten Ausführung dieser Maschine bis heute erhalten geblieben. Dieses verfügt zusätzlich zu den oben schon erwähnten Teilen über eine Vorrichtung zur Veränderung der schwingenden Länge des Rohrblattes. Dadurch konnte Kempelen Intonation imitieren, etwas das spätere Generationen sprechender Maschinen immer noch nicht richtig können werden. Dieses letzte Exemplar kann in München im Deutschen Museum besichtigt werden. Es ist dort in der Abteilung für Musikinstrumente ausgestellt, was bei der Art der Bedienung auch nicht ganz abwegig ist. Hartmut Traunmüller hatte 1997 an der originalen Maschine Funktionsversuche durchführen können. Er beschreibt seine Erfahrung so: *„Die Stimme war der eines Kindes ähnlich und ziemlich laut. Mehrere wesentliche Details des Gerätes*

[17] Kempelen, *Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine*; zitiert nach: Traunmüller, *Wolfgang von Kempelens sprechende Maschine*.

*waren aber nicht mehr in funktionsfähigem Zustand."*[15]

Am *Kempelen Farkas Speech Research Laboratory* des Institutes für Linguistik an der Akademie der Wissenschaften in Budapest wurde ein Nachbau der Maschine angefertigt. Es können Tonmitschnitte von Proben des Apparates aus dem Internet heruntergeladen werden.[18] Die Stimme ähnelt der eines Mädchens.

### 2.1.3 Euphonia, die erstaunliche sprechende Maschine

Erfinder der Euphonia ist Joseph Faber, ein Deutscher der schließlich nach Amerika auswanderte:

> *Faber, Joseph (Erfinder der Sprachmaschine, geb. zu Freiburg im Breisgau zu Anfang dieses Jahrhunderts; endete um 1850 in Amerika durch Selbstmord sein Leben).*[19]

Er verfiel nach einer schweren Krankheit in eine Hypochondrie, von der er sich durch mechansiche Arbeiten befreien wollte.

> *So verlegte er sich zuerst auf's Holzschnitzen, als ihm Kempelens Schrift* [...] *in die Hände kam und er auf die Idee verfiel, eine Sprachmaschine zu construieren.*[19]

Faber entwickelte auf Basis von Kempelens Buch eine im Funktionsumfang verbesserte Maschine. Herausragendes Merkmal war die Zunge, die es ihm ermöglichte mehr Laute wiederzugeben. Er nannte seine Maschine Euphonia.

> *Die Maschine hatte Aehnlichkeit mit einer kleinen Stubenorgel, besaß aber nur Eine Pfeife. Die Luft erhielt sie durch einen Blasebalg, den man mit den Füßen trat; die Veränderungen der Sprachlaute wurden aber durch 16 Tasten bewirkt. Die wesentlichen Theile der Stimm- und*

[15] Traunmüller, *Wolfgang von Kempelens sprechende Maschine*.
[18] URL: http://www.heise.de/ct/ftp/projekte/sprachsynthese/.
[19] Wurzbach, *Biographisches Lexikon des Kaiserthums Östereich*,S. 127f.

> *Sprachlaute waren größtentheils aus elastischem Gummi der Natur nachgebildet und die verschiedenartigsten Stellungen und Bewegungen derselben konnten durch Drähte hervorgebracht werden, welche sich auf mancherlei Weise an sie befestigten und durch Niederdrücken der Tasten bewegt werden konnten.* [sic][19]

Das interessante an dieser Maschine ist jedoch eher ihre Präsentation. So wurde an der Vorderseite der Maschine ein Gesicht einer Frau angebracht und der Körper durch Kleider angedeutet. Insofern könnte man die Euphonia als einen der ersten virtuellen Charaktere mit eingebauter (nicht programmierbarer) Sprachsynthese verstehen:

> *This was surmounted by a half-length weird figure, rather bigger than a full-grown man, with an automaton head and face looking more mysteriously vacant than such faceslook. Its mouth was large, and opened like the eyes of Gorgibuster in the pantomime, disclosing artificial gums, teeth, and all the organs of speech.*[...] *As a crowning display, the head sang a sepulchral version of "God save the Queen," which suggested inevitably, God save the inventor.*[...] *Sadder and wiser, I, and the few visitors, crept slowly from the place, leaving the professor with his one and only treasure - his child of infinite labour and unmeasurable sorrow.*[20]

Nachdem er auch in Amerika keinen Erfolg hatte und keine Beachtung fand, zerstörte er die Maschine und nahm sich das Leben.

## 2.2 Elektrische, analoge Sprachmaschinen

Schon kurz nach der Entdeckung des elektrischen Stroms erfand man den elektrischen Schwingkreis. Die Ingenieure sahen eine Analogie zwischen den elektrischen Schwingungen und den Schwingungen des Schalls. So wird diese Analogie zur Grundlage der Telephonie und des Sprechfunkes. Für

[19] Wurzbach, *Biographisches Lexikon des Kaiserthums Östereich*
[20] Hollingshead, *My Lifetime*; zitiert nach: Connor, *Euphonia*.

die Sprachsynthese ergeben sich zwei neue Ansätze, die beide auf dieser Analogie beruhen.

Bei der Vokaltraktsimulation wird der Vokaltrakt als Abfolge einzelner Filter interpretiert. Das Grundgeräusch passiert alle diese Filter nacheinander, am Ende kommt der Laut heraus. Man kann es sich in etwa so vorstellen, daß der Vokaltrakt in kurze scheibenförmige Abschnitte bestimmter Breite zerlegt wird. Jeder dieser Abschnitte moduliert das Tonsignal auf eine bestimmte Weise, dies wird durch einen Filter dargestellt. Es hat sich der Begriff *Leitungsanalogie (transmission line analog)* etabliert.

Bei der Formantsynthese *(terminal analog)* geht man einen anderen Ansatz, der quasi orthogonal zur Vokaltraktsimulation liegt. Hierbei wird der Laut als Mischung verschiedener Teiltöne angesehen, die durch Überlagerung zum gewünschten Ton verschmelzen. Dieser Ansatz ahmt nicht mehr den natürlichen Apparat nach, sondern versucht die Funktion in einer behaviouristischen Weise zu ermitteln. Die so ermittelte Funktion wird auch als Vokaltraktfunktion bezeichnet. Die einzelnen Töne werden durch jeweils eigene Filter aus einem Grundsignal erzeugt. Während bei der Vokaltraktsimulation alle Filter von einem Singal nacheinander durchlaufen werden, werden die Filter bei der Formantsynthese in der Regel parallel durchlaufen und die Teilsignale am Ende zu einem Gesamtsignal zusammengesetzt. Aus Untersuchungen erfuhr man, daß im Prinzip drei Formanten ausreichen um die Sprache erkennbar nachzuahmen.

Diese beiden Verfahren sind bis heute gängige Verfahren im Bereich der Forschung. Während diese Verfahren in der elektronischen Zeit auf die Filterfunktionen beschränkt sind, die mit den damals zur Verfügung stehenden elektrischen Bauteilen realisierbar waren, können heute beliebige Filter berechent werden. Die grundlegende Funktion ist aber immer noch dieselbe. Der Rechenaufwand ist bei der digitalen Simulation jedoch so hoch, daß diese Verfahren heute fast ausschließlich auf die Forschung und den akademischen Bereich beschränkt sind.

### 2.2.1 Die ersten Töne

Die Anfänge der elektrischen Sprachsynthese nehmen sich im Vergleich zu den damaligen mechanischen Lösungen recht einfach aus. So versucht man erst einmal die einfachen mechanischen Geräte in ihrer Funtkion nachzuahmen. Ein Beispiel hierfür ist die Veröffentlichung J. Q. Stewarts im Nature 110, 1922. In *„Electrical Analog of the Vokal Organs"* beschreibt er einen im wesentlichen aus drei ineinander verschachtelten Schwingkreisen aufgebauten Apparat, mit dem man kontinuierlich von einem Vokal zum anderen wechseln kann, und dabei sogar einige Diphthonge hervorzubringen vermag.[21]

### 2.2.2 Der Voder

In den 30'er Jahren des letzten Jahrhunderts forschten Homer Dudley und seine Mitarbeiter bei den Bell Laboratories an der Kompression von analog übertragenen Sprachsignalen. Ziel war es, die benötigte Bandbreite zur Übertragung von Gesprächen über die Ozeankabel zu reduzieren, und somit die teuren Verbindungen besser ausnutzen zu können. Hierbei entwickelten sie die Grundlagen für die Formantanalyse und die Formantsynthese. Das Gerät, das sie zu Demonstraionszwecken entwickelten, bestand aus zwei Komponenten, einem Gerät zur Analyse/Kompresion, dem Vocoder, und einem Gerät zur Synthese, dem Voder. Der Voder sollte durch Signale des Vocoders gesteuert werden.

> *In conceiving the vocoder, Dudley recognized the carrier nature of speech. He observed that the speech signal is formed by modulating (with the slowly changing vocal resonances) the spectral shape of the sound produced by vocal sources, The vocal sound sources may be periodic, as produced by vocal cord vibration, or aperiodic, as produced by turbulent airflow at a constriction.*
>
> *The modulations in shape of the speech spectrum could, therefore, be measured in terms of the relative energy in contiguous filter bands, and*

[21] Cater, *Electronical Speaking: Computer Speech Generation*; zitiert in: Rubin / Vatikiotis-Bateson, *J. Q. Stewart, Electrical analog of the vocal organs, Nature, 1922.*

*the periodic (voiced) or aperiodic (unvoiced) sources could be characterized by a "pitch" detector (a frequency meter). The signal could be reconstituted (synthesized) from these data by allowing to amplitude modulate the respective outputs of an identical filter bank which was excited by either a periodic pulse source or a noise source.*[22]

Auf diese Weise war eine Reduzierung der zur Übertragung von Sprache nötigen Bandbreite auf etwas über 300 Hz möglich, was gegenüber der damals gängigen Übertragungstechnik eine Reduzierung auf ein Zehntel entsprach.

Dudley und seine Mitarbeiter bemerkten, daß man den Voder, die Synthesekomponente, auch als eigenständiges über eine Tastatur zu steuerndes Gerät zur künstlichen Erzeugung von Sprache einsetzten konnte. Der interessante und unterhaltsame Charakter dieses Gerätes, das zeitgleich neue wissenschaftliche Prinzipien und Erkenntnisse sowie die technischen Möglichkeiten aufzeigen konnte, machten es zu einem idealen Exponat für die *Bell System Exhibit* auf der Weltausstellung 1939 in New York und San Francisco.

Obwohl für die Sprachübertragung gedacht, fanden die Erkenntnisse und Erfahrungen Dudleys und seines Teams vor allem bei der Verschlüsselung von vertraulichen Sprachsignalen während des zweiten Weltkrieges Verwendung. Die von ihm erarbeiteten Techniken sind bis heute grundlegender Bestandteil von Systemen zur Sprachübertragung.

### 2.2.3 Pattern Playback

In den späten 40'er Jahren des letzten Jahrhunderts bauten Franklin S. Cooper und seine Kollegen an den Haskins Laboratories die Pattern-Playback-Maschine (PPM). Dabei handelt es sich um die Umkehrung eines Sprach-Spektrogramm-Aufzeichners, eines Sonographen. Sonogramme geben, meist als Grauwert, die Schallintensität auf einem zweidimensionalen Diagramm an. Auf der einen Achse wird die Frequenz aufgetragen, auf der anderen die

[22] Millman, *A History of Engineering and Science in the Bell System*,S. 101f.

Zeit. Die PPM wandelt ein solches Sonogramm wieder in Ton um. Dieses Gerät ermöglichte es neben aufgenommenen natürlichen Sprachbeispielen auch eigene synthetische Samples zu erstellen. Dazu wurde mit weißer Farbe auf ein Zellulose-Azetat-Band das Sonogramm aufgezeichnet. Diese Maschine hat insbesondere bei der Erforschung der „Cues", also die für das Verstehen wesentlichen Bestandteile von Sprache wertvolle Dienste geleistet.

Die PPM ist eher in die Kategorie der Formantsynthese-Maschinen einzuordnen, obwohl dies nicht ganz eindeutig ist. Die genaue Einordnung hängt stark von der Betrachtungsweise ab.

Die PPM besteht aus einer sehr starken Lichtquelle, die ein dünnes Band Licht herausgibt. Dieses wird konzentrisch auf eine rotierende Scheibe geworfen. Auf dieser Scheibe sind ringförmig 50 Tonspuren angebracht, die ähnlich denen eines Tonfilmes 50 Grundfrequenzen erzeugen. Durch die konstante Umdrehung der Scheibe von 1800 U/min ergeben sich an den einzelnen Tonspuren Grundtöne mit Frequenzen zwischen 120 bis 6000 Hz in jeweils 120 Hz Abstand. Das dünne Lichtband hat nun nebeneinander verschiedene Frequenzen aufmoduliert bekommen. Man könnte sie auch als 50 einzelne, dicht nebeneinander verlaufende Lichtstrahlen auffassen. Diese Linien werden an einem Spiegel reflektiert, und durch das Sonogramm geworfen. Dieses ist auf ein Band geschrieben. An den Stellen wo das Zellulose-Acetat-Band durchlässiger ist wird der passende Teil des Lichtbandes (Tons) mehr durchgelassen, an den anderen weniger. Durch den photoelektrischen Wiederstand werden die einzelnen Teile des Lichtbandes zu einem gemeinsamen Signal zusammengesetzt, daß nach Verstärkung durch einen Lautsprecher ausgegeben wird. Normalerweise ist das Zellulose-Acetat-Band phototechnisch behandelt, so daß über die Lichtdurchlässigkeit die Stärke des jeweiligen Teiltones geregelt wird. Bei den Freihandzeichnungen muß das Gerät im Reflektionsmodus betrieben werden, d.h. über dem Band nimmt ein Sensor das durch die weiße Farbe reflektierte Licht war.

Das Gerät wurde bis Mitte der 70'er Jahre für Studien eingesetzt, und dient nun als Attraktion für die Besucher:

> *The device now resides in the basement of Haskins Laboratories, in New Haven, Connecticut, where it is often shown to our many visitors.*[23]

### 2.2.4 OVE und PAT

Etwa zeitgleich arbeiteten Gunnar Fant an der Königlichen Technischen Hochschule in Stockholm und Walter Lawrence an Geräten zur Formantsynthese. Gunnar Fants OVE, der Name entstand spontan während eines Interviews und wurde später als Orator Vorbis Electricus erklärt, bestand aus mehreren Schaltkreisen, die Formanten nicht nur parallel sondern auch sequenziell verarbeiteten. Die erste Stufe kann mittels eines zweidimensionalen Zeigearms, ähnlich einem Plotter, bedient werden, indem man auf einer Tafel den Zeiger auf den gewünschten Vokal einstellt. Mit dem OVE in seiner ersten Ausführung, es gab mehrere Nachfolger, unter anderem das moderne Infovox-System, konnte man so Vokale nachahmen. Der Zweck der Forschungen war ebenso wie bei Dudley die Bandbreitenreduzierung für Telefonübertragungen. Allerdings arbeitet Fant an der schwedischen Sprache, und verlegt sich schließlich ganz auf die synthetische Spracherzeugung.

Der PAT, Parametric Artificial Talker, arbeit mit drei parallelen Formantresonatoren. Auf einer bewegten Glasscheibe waren ähnlich dem Pattern Playback Signale in ihrem zeitlichen Verlauf aufgezeichnet, die die Frequenz der drei Formanten, die Stärke der Stimme, die Stärke des Rauschens und die Grundfrequenz regelten. Als Stimme wurde in beiden Systemen eine periodische Tonquelle verwendet, beim PAT war dies ein Summer, damit wurden stimmhafte Laute erzeugt. Zur Erzeugung von stimmlosen Lauten verwendete man aperiodisches Rauschen. Anders als der OVE konnte der PAT auch Konsonanten nachahmen.

Im Gegensatz zum Pattern Playback, bei dem das gesamte Spektrum durch die fünfzig einzelnen Abschnitte gesteuert wurde, und die Formanten sich

[23] Rubin / Goldstein, *The Pattern Playback*.

aus dem aufgezeichneten Sonogrammen ergaben, wurde bei diesen beiden Maschinen auf die gezielte Erzeugung von drei Formanten Wert gelegt. Sie werden daher auch als die ersten Formantsynthesegeräte im engeren Sinne bezeichnet.[24]

### 2.2.5 Electrical Vocal Tract

1950 veröffentlichte H. K. Dunn ein Modell des Vokaltraktes, mit dem er die Entstehung von Vokalen nachvollziehen kann. Das mathematische Modell wurde als elektrisches Gerät umgesetzt, bei dem 25 hintereinandergeschaltete Schwingkreise jeweils 5 mm lange Segmente von etwa 6 $cm^2$ Stirnfläche des Vokaltraktes nachbilden. Die gesamte Kette kann dann in zwei Teilsegmente zerlegt werden, zwischen die eine variable Spule eingesetzt wird. Diese simuliert den Zungenhügel. Eine weitere variable Spule am Ende stellt die Lippenstellung dar.[25]

## 2.3 Sprechende Computer

Mit der Verbreitung der modernen Rechentechnik veränderte sich auch die Sprachsynthese. Der Rechner wird auch als symbolverarbeitende Maschine bezeichnet, woraus sich eine neue Anwendung für die Sprachsynthese ergibt. War sie bisher im wesentlichen zu Forschungszwecken betrieben worden, kann sie nun auch zur Kommunikation zwischen dem Anwender und der Maschine beitragen. Insbesondere die Umwandlung von geschriebenen Text in Sprache stellt heute das Hauptanwendungsgebiet in der außerakademischen Verwendung. Solche Systeme werden Text-to-Speech, kurz TTS genannt. Der Rechner als Vorleser dient dabei unter anderem der automatischen Datenauskunft. Beispielsweise sei hier das Wetterauskunftsystem für den Flugverkehr VOLMET erwähnt, das für mehrere Flughäfen die Wetterbedingungen an die Piloten weiterreicht. Der VOLMET-Funk von Berlin, für die Flughäfen Berlins sowie Dresden, Leipzig, Warschau, Prag, Wien und

[24] Lemmetty, *History and Development of Speech Synthesis*.
[25] Rubin / Vatikiotis-Bateson, *Dunn's Electrical Vocal Tract*.

Kopenhagen, wird auf der Frequenz 128.400 MHz übertragen. Man kann ihn mit geeigneten Empfängern hören. Die jeweiligen Wetterdaten werden nach einem Standardschema von einem TTS vorgelesen. Tonbeispiele können aus dem Internet[26] abgerufen werden.

Außerdem ermöglichte der Rechner große Fortschritte in den schon existierenden Verfahren. So ist man nicht mehr auf solche Filterfunktionen beschränkt, die durch elektrische Bauteile darstellbar sind, sondern man kann alle prinzipiell berechenbaren Filterfunktionen einsetzen. Der Einsatz zeigt Schwächen in den zugrundeliegenden Modellen. So ist etwa der bisherige Grundsatz bei der Vokaltraktsimulation ungültig. Man kann den Vokaltrakt eben nicht in kleine voneinander unabhängige Teilstücke zerlegen. Es gibt sogar sehr komplexe Wechselwirkungen zwischen den einzelnen Abschnitten.

Schließlich entstanden zwei neue Formen der Sprachsynthese. Die konkatenative Synthese und die lineare Prädiktion. Einhergehend mit diesen neuen Verfahren trat eine Änderung an den Anspruch der Sprachsynthese ein. War es bisher Ziel, möglichst natürlich klingende Sprache vollständig künstlich hervorzubringen, so ist es nun das Ziel, ein ausreichend verständliches und als Sprache erkennbares Signal zu erzeugen. Der Anspruch ist vom gänzlich künstlichen Nachbau einer menschlichen Handlung weg zum pragmatischen Werkzeug zur Lösung von Aufgaben gewandert. War die Sprachsynthese bei Kempelen neben allen ideellen Ansichten noch Selbstzweck, so ist sie heute Mittel zum Zweck.

### 2.3.1 Text-zu-Sprache

Einige Menschen bezeichnen den Rechner gern als symbolverarbeitende Maschine. Darin drückt sich eine gewisse Auffassung von den Möglichkeiten des Einsatzes von Rechnern aus. Faßt man nun Sprachstücke als Symbole auf, ebenso wie die Buchstaben eines Textes, so liegt die Idee nahe, nach gewissen Regeln die Buchstabenfolge in eine Lautfolge umzurechnen, Also eine Symbolfolge in eine andere zu transformieren. Und genau dazu sind

[26] Internet: http://flugfunk.de, Menüpunkt VOLMET.

Computer ja nach dieser Auffassung gedacht.

TTS-Systeme sind die naheliegendste Anwendung der Sprachsynthese auf Rechnern. Diese Systeme sind um einiges komplexer als die vorigen Systeme. Die Ausgabe ist aber auch wesentlicher besser geworden. Im Grunde besteht ein TTS-System aus zwei Komponenten, die ihrerseits wieder in beliebig viele Unterkomponenten gegliedert werden können. Je nach Aufwand erhält man dann unterschiedliche Qualitäten. Im einfachsten Fall besteht ein TTS aus einem Text-zu-Phonem-Übersetzter und einem Phonem-zu-Ton-Übersetzer. Der Text-zu-Phonem-Übersetzer kann mehrere Stufen vorgschaltet bekommen, die die Texte normieren, zum Beispiel Abkürzungen in ausgeschriebene Worte umwandeln. Dem Phonem-zu-Ton-Übersetzer können Stufen als Filter nachgeschaltet werden, z.B. zur Steuerung der Intonation. Bei den komplexeren Systemen entsteht das Problem, das Stufen Daten zur Steuerung anderer Stufen durchreichen müssen. Dazu werden besondere Symbolfolgen in den Datenstrom eingefügt.

### 2.3.2 1+1=zwei

Es gibt bekanntlich nichts, was Unix nicht kann. So verwundert es denn auch nicht, daß bereits frühe Versionen Funktionen zur Sprachsynthese besitzen:

> *In 1972, the standard Unix manual (3rd edition) included commands to process text to speech, form text analysis, prosodic prediction, phoneme generation, and waveform synthesis through a specialized piece of hardware. Of course Unix had only about 16 installations at the time and most, perhaps even all, were located in Bell Labs at Murray Hill.*[27]

### 2.3.3 DECtalk

In den 80er Jahren entwickelte Klatt das DECtalk genannte System, ein Echtzeit Sprachsynthesesystem in Hardware. Die Hardwarelösung war notwen-

[27] festvox.org, *Overview of Speech Synthesis*.

dig, da die gängigen allgemeinen Rechnersysteme nicht über die nötige Rechenkraft verfügten.

> *A laboratory text-to-speech system, or a development system, is best implemented on a large general-purpose digital computer. The flexibility and nearly unlimited computational resources outweigh disadvantages of non-real-time output. However, practical commercial systems must realize real-time operation at a reasonable cost/ performance tradeoff, while simultaneously providing additional features such as a flexible user interface and telephonics for many commercial applications.*[28]

Beim DECtalk handelt es sich im Prinzip um das in Hardware „gegossene" KlattTalk, ein TTS-Software-System. Es sollte für etwa 1000,- USD an sprechbehinderte Menschen verkauft werden. Während der Entwicklung stellte sich dann heraus, daß der Verkaufspreis aber um den Faktor vier höher liegen würde, womit das Gerät für die angestrebte Zielgruppe nicht mehr erschwinglich war. Das Gerät basierte hardwareseitig auf einem Motorala 68000 General Purpose Prozessor, einem Texas Instrument TMS-32010 Signalprozessor für die Formantsynthese und zur Steuerung verschiedener Parameter für die Soundausgabe sowie Arbeitsspeicher in dem unter anderem ein 6000 Worte umfassender Bestand an Ausnahmen gespeichert ist.

### 2.3.4 Speak'n'Spell

In den siebziger Jahren des letzten Jahrhunderts wurde Sprachsynthese nicht nur besser, sie wurde auch preiswerter. Ein sehr schönes Beispiel dafür ist das Lernspielzeug Speak'n'Spell von Texas Instruments. 1978 begann der Verkauf des Gerätes, es wurde in mehreren Versionen mit einigen kleinen Änderungen verkauft und war eines von drei Lernspielzeugen aus dem Hause. Man kann insgesamt fünf Sprech- und Buchstabierspiele damit spielen.

[28] Klatt, *Review of text-to-speech conversion for English*, S. 775.

> *DALLAS (June 11, 1978) - A new speech synthesis monolithic integrated circuit has been developed by Texas Instruments Incorporated. It marks the first time the human vocal tract has been electronically duplicated on a single chip of silicon.*[29]

Der verwendete Chip setzt die etwa zur gleichen Zeit modern werdende lineare Prädiktiktion ein. Der Chip zur Synthese verwendet 12 Parameter, 10 Filterkoeffizienten sowie Höhe und Stärke, als Eingaben. Diese sind in einem ROM abgspeichert und werden von einem Schaltkreis auf dem Chip entschlüsselt. Sie stellen dann die zeitveränderliche Beschreibung des linear prädiktiven Modells dar. Als Eingänge in die Filter dienen schon wie bei OVE ein periodisches Signal und ein aperiodisches Rauschen. Damit werden stimmhafte und stimmlose Laute erzeugt. Dazu ist der Chip mit zwei separaten Logikblöcken ausgestattet.

Es wird eine lineare Prädiktion 10. Ordnung verwendet, was durch ein zehnstufiges Filtergatter realisiert wird. Der 8kHz-Chip kann bis zu 10000 Sprachstücke pro Sekunde berechnen. Er kann Sprache aus gespeicherten oder aus übertragenen (Online-) Daten erzeugen.[29] Das Gerät konnte in seinem Sprachumfang durch Zusatzmodule erweitert werden, für die am Batteriefach ein Kasetteneinschub vorhanden war.

Das Gerät war insgesamt sehr erfolgreich und wurde auch in anderen Ländern verkauft, so auch in Deutschland unter dem Namen *buddy - Der sprechende Lerncomputer*.[30]

[29] Texas Instrument Incorporated, *First Single-Chip Speech Synthesizer*, Pressemitteilung.
[30] www.99er.net, *The Texas Instruments Speak & Spell*.

### 2.3.5 Lineare Prädiktion

Die lineare Prädiktion ist eigentlich ein Verfahren in der Signalverarbeitung zur Datenkompression. Dabei wird angenommen, daß das erwartete Signal als lineare Kombination vorhergehenden Signale aufgefasst werden kann. Das Signal wird dabei als Summe der vorausgegangen Signale aufgefasst, wobei diese mit einem Koeffizienten modifiziert werden.

$$\hat{s_n} = q_1 s_{n+1} + q_2 s_{n+2} + ... + q_m s_{n+m}$$

Dabei wird die Anzahl der rückwärtig betrachteten Signale als Stufigkeit angegeben. Ein Verfahren, wie bei dem Speak'n'Spell von Texas Instruments, das 10 zurückliegende Signale verwendet wird als 10-stufige Lineare Prädiktion bezeichnet (LP10).

Zurückliegende oder vorrausgehende Signale meint in diesem Zusammenhang von der Position des zu berechnenden Signales aus gesehen. Dies kann vom Zeitpunkt des Betrachters aus gesehen durchaus noch in der Zukunft sein.

Häufig werden konkatenative Systeme, die nicht die Segmente als reinen Sample abgespeichert haben, sondern lineare Prädiktion zur Datenreduzierung verwenden, aber auch zur Gestaltung der Übergänge zwischen den Lauten, von den konkatenativen abgegrenzt und als linear prädiktive Systeme bezeichnet. Streng genommen sind sie dadurch aber immer noch konkatenative Systeme.

### 2.3.6 Konkatenative Sprachsynthese

Bei der konkatenativen Sprachsynthese handelt es sich um ein Verfahren, bei dem Sprachsegmente nach Bedarf aneinandegehängt werden. Die ersten Versuche basierten auf den Phonemen. Der Text lag als Folge von Phonemen vor. Das System mußte nur die zu den Phonemen gehörenden Tonsegmente in der vorgebenen Reihenfolge aneinanderhängen. Prinzipiell ist dieses Verfahren recht einfach und kommt mit niedrigem Speicherbedarf aus, da jedes Phonem genau einmal gespeichert wird. Es zeigte sich aber

recht schnell, daß mit diesem Verfahren keine Wörter gesprochen ausgegeben werden können, die man als zusammenhängend wahrnimmt. Dies liegt an den Übergängen zwischen den Lauten in der natürlichen Sprache. Die Aussprache eines Lautes hängt von seinem Vorgänger und von seinem Nachfolger ab. Da diese Abhängigkeit nicht berücksicht ist, klingen die Ausgaben abgehackt und unzusammenhängend.

Die nächste Idee war es, nicht die Laute selbst als Segmente zu verwenden, sondern die Übergänge zwischen den Lauten. Diese nennt man Diphone. Dabei handelt es sich um den Teil zwischen dem Invarianten Teil des ersten Phonems bis zum Invarianten Teil des zweiten Phonems. Damit war es nun möglich eine fließendere Sprachausgabe zu erzeugen. Gleichzeitig quadrierte sich aber der Speicheraufwand, da nun für alle Paarungen von Lauten die Tonsegmente gespeichert werden mußten. Die Ausgabe ist wesentlich besser und kann mit relativ geringem Rechenaufwand bewerkstelligt werden. Es zeigten sich jedoch auch hier Grenzen, da nicht nur der direkte Vorgänger, sondern manchmal auch dessen Vorgänger, die Aussprache eines Lautes beeinflussen. Es wären also Triphone nötig. Man kann dies in einigen Fällen sogar zu Quadrophonen u. ä. erweitern. Dies ist jedoch nicht mehr praktibel. Der Speicheraufwand für alle möglichen Kombinationen wäre schlicht zu groß geworden, um ihn noch bewältigen zu können. Man schuf also eine Lösung die auf den Diphonen aufbaut, aber auch längere Passagen, manchmal sogar ganze Redewendungen enthält. Diese liegen in Form von Wörterbüchern vor. Da man anfänglich nur häufig vorkommende Worte abspeicherte, die durch die Verkettung von Diphonen nicht adäquat wiedergegeben wurden, nannte man diese Wortlisten Ausnahmewörterbücher (Exception Dictionaries). Dieser Name hat sich bis in die heutige Zeit erhalten, obwohl es mittlerweile eher die Ausnahme ist, ein Wort nicht im Wörterbuch zu finden. Hierbei bezieht sich der Begriff Wort nicht auf das grammatikalische Wort allein, sondern umfasst häufige Lautfolgen, meist Wortstämme, und gelegentlich auch ganze Wortgruppen. Als Eingabe dient nun nicht mehr eine Phonemschrift, sondern eine Art Segmentschrift, die angibt welche Einträge des Wörterbuches abgespielt werden sollen. Die Diphone dienen dabei als Kitt, um die größeren Segmente zu verbinden, oder

mit Endungen, Vorsilben u.ä. zu versehen. Obwohl es keine Phonemschrift mehr ist, wird sie in der Literatur jedoch häufig noch als solche bezeichnet, wohl Mangels eines besseren Begriffes.

Durch diese Veränderung wurden konkatenative Sprachsynthesesysteme zu gut klingenden und verbreiteten Systemen. Es sind aber auch die Systeme mit dem geringsten Entwicklungspotential, denn hier steckt die eigentliche Arbeit in der Aufstellung der Diphon-Datenbanken (die auch die Wörterbücher beinhalten). Hinzu kommt der Nachteil, daß sich keine gute Form der Betonung darstellen lässt, da die einzelnen Diphone genau aufeinander abgestimmt sind, und eine Tonhöhenänderung in einer Silbe eines Wortes dieses Wort auseinanderfallen lassen würde. Man versucht durch Tonhöhenänderung für ganze Worte dieses Problem zu umgehen. Eine andere Lösung besteht darin, hinter die Sprachausgabe des Systems ein zweites System zu schalten, daß die Tonhöhenänderung berechnet. Diese Art fällt dann in die Klasse der hybriden Systeme, da hier zwei Systeme, konkatenativ und entweder formantsynthetisch oder vokaltrakt-simulierend, kombiniert werden.

Ein bekanntes Beispiel für konkatenative Systeme ist das freie Festival-TTS-System, und viele seiner Ableger, wie etwa MBROLA.

# 3 Entwicklungen

Die Vokaltraktsynthese dürfte, obwohl ihre Ergebnisse bisher eher enttäuschend sind, diejenige mit dem höchsten Entwicklungspotential sein. Hier besteht das Problem zur Zeit vor allem in der hohen mathematischen Komplexität der Modelle, der geringen Kenntnis über die Abläufe am echten Vokaltrakt und der noch zu geringen Rechenleistung. Diese Probleme dürften sich jedoch in den nächsten Jahrzehnten deutlich entschärfen, so daß hier mit großen Fortschritten gerechnet werden kann. Prinzipiell erlaubt diese Herangehensweise die Synthese aller Aspekte der Sprache.

Die Formantsynthese ist ein reichlich eingesetzes Verfahren, daß nach wie vor als Bestandteil der Formantanalyse-Formantresynthese zur Datenkompression von Sprachsignalen Anwendung findet. Die Konkatenative Synthese ist ebenfalls eine weitverbreitete Technik, die durch ihren geringen Rechenaufwand besticht. Formantsynthese und konkatenative Synthese zusammen, in Form sogenannter Hybride, dürften in naher Zukunft den Hauptteil der kommerziellen Systeme ausmachen, da sich hier durch die Verbindung der beiden Technologien enorme Fortschritte abzeichnen. Solange die Vokaltraktsimulation diese Fortschritte nicht auf- und überholt, werden Hybride den Standard darstellen.

Neben der Synthesetechnik an sich ist die Aufbereitung der Sprache vor der Synthese zunehmend bedeutsam. TTS-Systeme sind im praktischen Einsatz heute der am weitesten verbreitete Anwendungsfall. Jedoch muß die Schriftsprache, in der die Texte in aller Regel vorliegen, erst in eine Phonem- bzw Segmentschrift umgewandelt werden. Die Regeln dazu sind recht komplex und noch nicht vollständig erfaßt. Hinzu kommt, daß für eine solche Umwandlung der Text nur aus vollständig aufgeschriebenen Wörtern bestehen darf. Normaler Text besteht aber eben auch aus Abkürzungen, Ver-

kürzungen, Zahlen und besonderen Symbolen. Diese müssen umgewandelt werden; ähnlich wie bei der Umwandlung von Makros in Programmiersprachen spricht man hier vom *Preprocessing*.

Es gibt aber auch andere Verfahren, um die Sprachausgabe zu verbessern. Bisher wurde TTS auf normalen schriftlichen Texten angewand. Seit einigen Jahren gibt es die Diskussion, bei Anwendungen die von vornherein Sprachausgabe verwenden, besondere Notationen zur Steuerung der Betonung, der Aussprache u. ä. zu verwenden. Hierbei wird wie im *World Wide Web* eine Markup-Language, eine Markierungssprache verwendet. Dies hat den entscheidenden Vorteil, daß weitere Informationen zugesteuert werden können. Die Texte müssen natürlich dementsprechend aufbereitet werden. Die einheitliche Markup-Language steht noch aus, da die Anforderungen noch nicht vollständig bekannt sind, und erst noch erprobt werden müssen.[31]

[31] Festival Handbuch, *XML / SGML mark-up*.

# 4 Anwendungen

Bei der Sprachsynthese gibt es schon seit ihren Anfängen den Wunsch sie praktisch einzusetzen. Wenn man von Systemen mit begrenztem Vokabular absieht, also solchen, wo Satzstücke fertig aufgenommen wurden und nach Bedarf aneinandergereiht werden, z. B. Bahnhofsansagen, dürfte der bisher größte und wohl auch erfolgreichste Anwendungsfall die Unterhaltung sein. Wolfgang von Kempelen führte seine sprechende Maschine zusammen mit dem Schachspielenden Türken dem Publikum vor. Allerdings hielt sich die Begeisterung in Bezug auf die Maschine stark in Grenzen, vieleicht wegen dem sehr viel interessanteren und berühmteren Schachautomaten.

Joseph Faber versuchte ebenfalls mit seiner Euphonia als Unterhalter aufzutreten, was ihm keinen Erfolg einbrachte und ihn schließlich innerlich zerbrechen ließ.

Der Voder mußte auf der Weltausstellung 1939 in New York und San Fransico als Publikumsköder Leute auf den Bell-Stand ziehen. Außerdem sollte er als Technologie- und Leistungsdemonstration herhalten.

In der heutigen Zeit hat vor allem die Computerspieleindustrie den Wert von computergenerierter Sprache erkannt, vor allem in Verbindung mit virtuellen Charakteren. So verwendet das verbreitete Multiplayerspiel *Unreal Tournament 2004* von *Epic Games* TTS um Textnachrichten anderer Spieler vorzulesen. Die Ausgabe ist zwar nicht überragend, aber einige Spieler sehen es als Vorteil, da sie nun nicht mehr lesen müssen, und ihre Augen somit auf das Geschehen des Spiels lenken können. Nebenbei sei erwähnt, daß dieses Spiel auch das Gegenstück zur Sprachsynthese, die Spracherkennung, verwendet, um Befehle an Bots, so heißen die virtuellen Charaktere die als Mitstreiter agieren, erteilen zu können. Dabei beschränkt sich diese

Technik aber auf vordefinierte Schlüsselworte.[32]/[33]

Neben der Unterhaltung versucht man auch, sprachbehinderten Menschen die Möglichkeit zu geben, sich sprachlich verständlich zu machen. Wohl berühmtestes Beispiel ist der Sprachcomputer des Physikers Stephen Hawking. Stephen Hawking erkrankte an ALS und hatte ab 1985 nach einer Operation keine Möglichkeit mehr, selbst zu sprechen. Walt Woltosz hörte von dem Leiden und gab Stephen Hawkin ein Programm, das er Equalizer nannte. Dieses Programm ermöglicht es Stephen Hawking mittels eines Zeigegerätes aus einer Menüstruktur Wörter zu Sätzen zu kombinieren, und schickt diese dann an ein TTS-System.

> *I have also given many scientific and popular talks. They have all been well received. I think that is in a large part due to the quality of the speech synthesiser, which is made by Speech Plus. One's voice is very important. If you have a slurred voice, people are likely to treat you as mentally deficient: Does he take sugar? This synthesiser is by far the best I have heard, because it varies the intonation, and doesn't speak like a Dalek. The only trouble is that it gives me an American accent.*[34]

Weiterhin gibt es die Möglichkeit Sehbehinderten Menschen die Kommunikation mit Rechnern zu vereinfachen, indem man sich der Sprachsynthese bedient. So kann man Bildschirminhalte vorlesen lassen. In der Regel wird diese Technik in Kombination mit einem Braille-Display oder der Großschriftdarstellung verwendet. Hierbei finden aber vor allem die billigeren einfacheren Synthesetechniken Vorliebe, da es den Benutzern weniger um die Natürlichkeit der Stimme als mehr um die Effizienz der Schnittstelle geht:

> *Zudem sind mir die hochwertigen Sprachausgaben zu träge, weshalb ich und auch viele Kunden mit den "schlechten Sprachausgaben" vorlieb nehmen. Damit kann man einfach schneller arbeiten.*[35]

[32] www.unrealtournament.com, *UT2004 Features*.
[33] www.driverheaven.net, *Unreal Tournament 2004*.
[34] Hawkin, *Disabilty – my experience with ALS*.
[35] Jaklin / www.faz.net, *Für Sehende muss es das teure TTS-System sein*.

Ein drittes Anwendungsfeld ist die sprachliche Ausgabe generierter Daten zu Informations- und Alarmierungszwecken. Ein Beispiel für einen Informationsdienst ist das weiter oben schon beschriebene VOLMET-System, bei dem Wetterdaten über einen Sprechfunkkanal an die Piloten übertragen werden. Hier lösen TTS-Systeme zur Zeit die Bandansagen ab, da Bandansagen in regelmäßigen Abständen neu erstellt werden müssen, und somit einen hohen Personalbedarf mit sich bringen. VOLMET ist zur Zeit aber gleichzeitig eine Ausnahme. Im Allgemeinen wird bei Informationsdiensten, die mit einem beschränktem Vokabular arbeiten, die Erzeugung aus vorgefertigten Textbausteinen bevorzugt, da sie gegenüber der vollständigen Synthese eine wesentlich natürlichere Ausgabe erzeugt. Bei VOLMET zählt aber nicht das Merkmal der Natürlichkeit, sondern der Verständlichkeit und Unmißverständlichkeit.

Ein Beispiel für den Einsatz von TTS zu Alarmierungszwecken bietet das Rechenzentrum der Universität Hohenheim.

> *Die Sache mit dem TTS-System (Text To Speach) war zunächst eine Spielerei, bis es zum ersten mal passierte... Eine AFS-Datenbank lief nicht und keiner hat's gemerkt und keiner hat's gesagt... keiner, außer dem Skript das da 'zufällig' lief.*[36]

Ein Skript zur Überwachung von Rechnern erzeugt im Fehlerfall neben der „normalen" Bildschirmausschrift eine gesprochene Fehlerausgabe. Dabei wird für die Sprachausgabe eine völlig andere Formulierung als für die Bildschirmausschrift verwendet, daß System ist also streng genommen kein TTS, sondern es bedient sich einer TTS-Software. Dieses Neuformulieren wird auf die Hörgewohnheiten zurückgeführt, und leicht ironisch auf den Punkt gebracht: *„Fazit: Ellenlange und reichlich interpunktierte Fehlermeldungen für die Sprachausgabe!"*[36]

Über die drei bereits genannten Anwendungsbereiche hinaus sieht es zur Zeit düster aus. Obwohl die Anbieter von Sprachsynthesesystemen nicht müde werden, die Vorteile der Technologie immer wieder zu betonen, bleiben die wirtschaftlichen Erfolge, der erwartete Durchbruch aus. Dies liegt

[36] Feiler, *Afs-Server Überwachung an der Universität Hohenheim*.

an unterschiedlichen Eigenschaften der Technologie, aber auch der Überbewertung der Sprache als Kommunikationsmittel selbst. So ist für viele Menschen die Verwendung von Vorleseprogrammen einfach unsinnig, da sie Texte sehr viel schneller lesen können als sie hören zu können. Hinzu kommt, daß spezielle Lesetechniken wie das Überfliegen oder Diagonallesen keine Entsprechung in der Sprache haben.

Andere meinen, die Technik sei bereits ausreichend gut, um in den sinnvollen Anwendungen eingesetzt zu werden, es würde aber für einen großen kommerziellen Durchbruch nicht reichen. Es gäbe demnach einfach nicht genug sinnvoller nutzbringender Anwendungen um eine große Allgemeintechnologie zu werden.

> *Die Crux der TTS Vermarktung sei kein Problem mangelnder technischer Reife, sondern das fehlender Anwendungsmöglichkeiten.*[37]

Ebenfalls gegen einen breiten Einsatz sprechen die Unzulänglichkeiten der Technologie. So stellen die Intonation und die Stimme, vor allem die Individualisierbarkeit der Stimme, die großen Probleme dar. Einige werden nie gänzlich gelöst werden, da hier die Grenzen des Berechenbaren erreicht werden.

> *Beim heutigen Stand der Technik ist die Grenze der erzielbaren Verständlichkeit und Natürlichkeit synthetischer Sprache kaum noch durch Faktoren technischer Art, sondern vielmehr durch unser begrenztes Wissen über die Akustik und Perzeption der Sprache gegeben. In der Forschung kann Sprachsynthese verwendet werden, um dieses Wissen zu testen. Es gibt jetzt automatische Methoden zur akustischen Analyse und Wiedersynthese von Sprache. Man kann dabei vor der Wiedersynthese gewisse Eingriffe machen, und z.B. versuchen, das scheinbare Alter des Sprechers zu verändern. Der Erfolg hängt davon ab, wie gut man die dabei wesentlichen Faktoren kennt.*[15]

[37] www.faz.net, *Es fehlt an Anwendungsmöglichkeiten.*
[15] Traunmüller, *Wolfgang von Kempelens sprechende Maschine.*

# 5 Quellenverzeichnis

Black, Alan W. / Lenzo, Kevin A.: *festvox.org: Overview of Speech Synthesis*. Internet: http://festvox.org/bsv/bsv-intro-ch.html [3.5.2005].

Black, Alan W. / Taylor, Paul / Caley, Richard:
*Festival Handbuch: XML / SGML mark-up*.
Internet: http://www.cstr.ed.ac.uk/projects/festival/manual/festival_10.html#SEC31 [19.5.2005].

Cater, John P.: *Electronically Speaking: Computer Speech Generation*. Howard M. Sams & Co, 1983.

Ciba, Anne: *Funktionales Stimmtraining 1*.
Internet: http://www.stimmbildung.com/funktionale-stimmpaedagigik.html [22.4.2005].

Connor, Steven: *Euphonia*.
Internet: http://www.dumbstruck.org/archive/euphonia.html [29.4.2005].

Feiler, Mathias: *Afs-Server Überwachung an der Universität Hohenheim*.
Internet: http://www.rz.uni-hohenheim.de/feiler/uni/afs/tools/ [24.5.2005].

Gósy, Mária: *On the Early History of Hungarian Speech Research*. In: International Journal of Speech Technology, 2000, 3, S. 155-164.

Hawkin, Stephen: *Disability– my experiance with ALS*.
Internet: http://www.hawking.org.uk/disable/dindex.html [24.5.2005].

Hollingshead, John: *My Lifetime*. London: Sampson, Low, Marston & Co, 1895, S. 67-69.

Jaklin, Manfred / Pretzer, Cornelia: *Für Sehende muss es das teure TTS-System sein*.
Internet: http://www.faz.net/s/RubCD175863466D41BB9A6A93D460B81174/Doc~E6414B294995647FAB593071BDAEBA669~ATpl~Ecommon~Scontent.html [24.5.2005].

Kempelen, Wolfgang von: *Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine*. J.V.Degen, 1791.

Klatt, Dennis H.: *Review of text-to-speech conversion for English*. In: Journal of the Acoustical Society of America, 1987, 82(3), S. 737-793.

Lemmetty, Sami: *History and Development of Speech Synthesis*.
Internet: http://www.acoustics.hut.fi/~slemmett/dippa/chap2.html [2.5.2005].

Lindsay, David: *Talking Head*. In: American Heritage on Invention & Technology, 1997, Summer, S. 57-63.

Millman, S. (Herausgeber): *A History of Engineering and Science in the Bell System*. AT&T Bell Laboratories, 1984, S. 99, 101-102.

Polivka, Rich: *The Texas Instruments Speak & Spell*.
Internet: http://www.99er.net/spkspell.html [18.5.2004].

Rubin, Philip / Goldstein Louis: *The Pattern Playback*.
Internet: http://www.haskins.yale.edu/haskins/MISC/PP/pp.html [29.4.2005].

Rubin, Philip / Vatikiotis-Bateson, Eric: *Simulacra*.
Internet: http://www.haskins.yale.edu/haskins/HEADS/simulacra.html [21.4.2005].

Rubin, Philip / Vatikiotis-Bateson, Eric: *Dunn's Electrical Vocal Tract*.
Internet: http://www.haskins.yale.edu/haskins/HEADS/SIMULACRA/dunn.html [18.4.2005].

Rubin, Philip / Vatikiotis-Bateson, Eric: *J. Q. Stewart, Electrical analog of the vocal organs, Nature, 1922*.
Internet: http://www.haskins.yale.edu/haskins/HEADS/SIMULACRA/stewart.html [21.4.2005].

Rubin, Philip / Vatikiotis-Bateson, Eric: *Kratzenstein*.
Internet: http://www.haskins.yale.edu/haskins/HEADS/SIMULACRA/kratzenstein.html [21.4.2005].

Strouhal, Ernst: *Kempelens Türke: Eine Schach-Metaphern-Maschine aus dem Spätbarrock*.
Internet: http://www.karlonline.org/402_5.htm [21.4.2005].

Traunmüller, Hartmut: *Wolfgang von Kempelens sprechende Maschine*.
Internet: http://www.ling.su.se/staff/hartmut/kempln.htm [22.4.2005].

Witzler, Ralf *Es fehlt an Anwendungsmöglichkeiten*.
Internet: http://www.faz.net/s/RubCD175863466D41BB9A6A93D460B81174/Doc~E39B7EC4739F44B0CA1F840D3C7CAD373~ATpl~Ecommon~Scontent.html [24.5.2005]

Wurzbach, Konstantin: *Biographisches Lexikon des Kaiserthums Österreich, Band 4: Egervari – Fürchs*. 1858. S. 124, 125.

Texas Instruments Incorporated: *First Single-Chip Speech Synthesizer*. Pressemitteilung: Dallas, 1978.
Internet: http://www.ti.com/corp/docs/company/history/pmos.shtml [18.5.2005].

Internetseiten mit nicht benannten Autoren:

de.wikipedia.org: *Diphon*.
Internet: http://de.wikipedia.org/wiki/Diphon [18.4.2005].

de.wikipedia.org: *Diphthong*.
Internet: http://de.wikipedia.org/wiki/Diphthong [18.4.2005].

de.wikipedia.org: *Formant*.
Internet: http://de.wikipedia.org/wiki/Formant [22.4.2005].

de.wikipedia.org: *Frikativ*.
Internet: http://de.wikipedia.org/wiki/Frikativ [22.4.2005].

de.wikipedia.org: *Konsonant*.
Internet: http://de.wikipedia.org/wiki/Konsonant [28.5.2005].

de.wikipedia.org: *Phon*.
Internet: http://de.wikipedia.org/wiki/Phon_%28Phon%29
[21.4.2005].

de.wikipedia.org: *Phonem*.
Internet: http://de.wikipedia.org/wiki/Phonem [21.4.2005].

de.wikipedia.org: *Phonetik*.
Internet: http://de.wikipedia.org/wiki/Phonetik [21.4.2005].

de.wikipedia.org: *Vokal*.
Internet: http://de.wikipedia.org/wiki/Vokal [28.5.2005].

flugfunk.de: *VOLMET*.
Internet: http://flugfunk.de Menüpunkt: VOLMET [3.5.2005].

lexikon.eventax.de: *Wolfgang von Kempelen*.
Internet: http://lexikon.eventax.de/wolfgang_von_kempelen/
[22.4.2005].

www.driverheaven.net: *Unreal Tournament 2004*.
Internet: http://www.driverheaven.net/articles/ut2004/ [24.5.2005].

www.heise.de: *Computertalk – Der Weg zur menschlichen Robovoice*.
Internet: http://www.heise.de/ct/ftp/projekte/sprachsynthese/
[25.4.2005].

www.unrealtournament.com: *UT2004 Features*.
Internet: http://www.unrealtournament.com/ut2004/features.php
[24.5.2005].