10

Hartmut R. Pfitzinger

Sprache aus dem Synthesizer

hpt@phonetik.uni-muenchen.de

Gleich zu Beginn dieses Beitrags zur KlangForschung möchte ich zwei möglichen Mißverständnissen vorbeugen und damit gleichzeitig skizzieren, worum es hier gehen wird:

1. Im Zentrum dieses Beitrags steht das Erzeugen von menschlichen Sprachlauten und verständlichen Äußerungen durch einen handelsüblichen, für Musiker bestimmten Synthesizer. Natürlich gehe ich auch auf die Sprachsynthese durch speziell hierfür entwickelte Synthesesysteme ein. Diese hochspezialisierten Systeme sind aber im Gegensatz zu Musik-Synthesizern trivialerweise geeignet, sprachliche Äußerungen zu erzeugen, während Musik-Synthesizer in erster Linie dazu gedacht sind, Instrumentalklänge und Geräusche zu generieren. Um mit für Musiker bestimmten Synthesizern verständliche Sprache zu erzeugen, müssen wir die Geräte bis zu einem gewissen Grad zweckentfremden.
2. Ich werde nicht darüber berichten, wie aus geschriebenen Texten Steuerparameter für Synthesizer abgeleitet werden, sondern wie ich statt dessen durch die akustische Analyse von sprachlichen Äußerungen zu einer sinnvollen manuellen Wahl der Parameter komme. Das dabei benötigte Hintergrundwissen über die akustische Struktur der menschlichen Sprache wird deshalb ebenfalls Teil dieses Beitrags sein.

Es geht also um die adäquate Einstellung derjenigen Klangparameter, die ein Musik-Synthesizer dem Klang-Programmierer anbietet. Die Funktion und Einstellung der Parameter wird im folgenden vorgeführt anhand des *Access Virus*, eines virtuell analogen Synthesizers, den Christoph Kemper, der Entwickler des Gerätes und zugleich Geschäftsführer von *Access*, dem Institut für Phonetik und Sprachliche Kommunikation zur Verfügung gestellt hat (siehe Beitrag 9 in diesem Band).

Dieses Gerät wird „virtuell analog“ genannt, weil es mit Hilfe eines digitalen Signalprozessors die Funktionsweise eines analogen Synthesizers simuliert. Durch seine spezielle interne Verschaltung der klangbestimmenden Einheiten ist dieser Synthesizer zwar besonders gut geeignet, auch Sprache zu synthetisieren, dennoch sollte jeder analoge Synthesizer mit Resonanzfiltern in der Lage sein, die hier vorgestellten „Sprachklänge“ zu erzeugen.

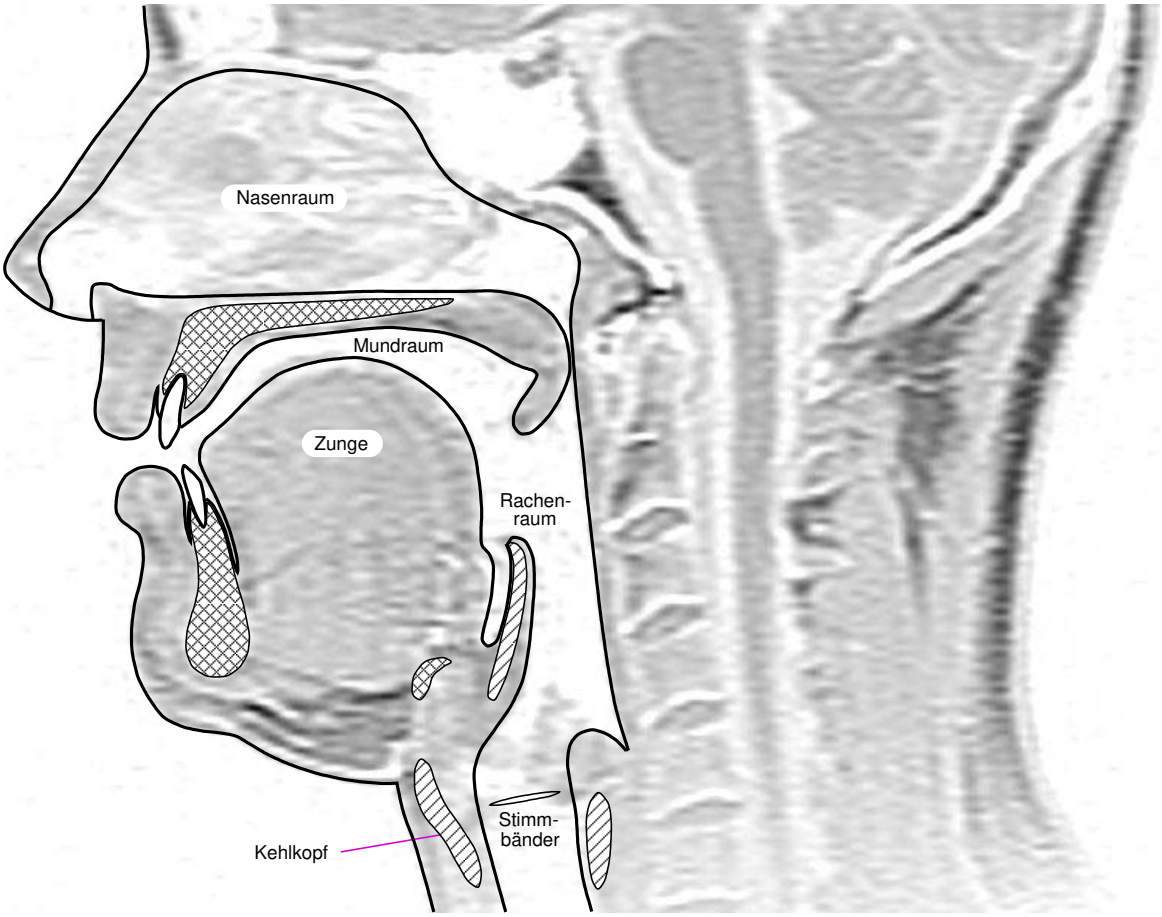

Abb. 10.1: Sagitalschnitt mittels Kernspintomographie durch den menschlichen Sprechapparat.

10.1 Das Quelle-Filter-Modell der Sprachproduktion

Um erklären zu können, wie die Akustik der gesprochenen Sprache strukturiert ist, muß man sich im klaren darüber sein, wie der Mensch Sprache hervorbringt. Der menschliche Sprechapparat ist in Abb. 10.1 als Sagitalschnitt dargestellt und läßt sich in zwei wesentliche Teile untergliedern:

1. den Kehlkopf, der durch die quasi-periodischen Schwingungen seiner Stimmbänder obertonreiche Töne erzeugen oder durch eine Engebildung Turbulenzen und damit ein Rauschen hervorbringen kann, (*Quelle*)
2. und den Mund-Nasen-Rachen-Raum, nachfolgend Ansatzrohr genannt, das wie jeder Hohlraum Resonanzeigenschaften besitzt und dadurch ein Formantfilter darstellt (*Filter*).

In der Phonetik wird ein Klangerzeugungsmodell, das diese beiden Einheiten umfaßt, *Quelle-Filter-Modell* genannt (siehe Abb. 10.2). In diesem Modell dient als Schallquelle entweder eine periodische Folge von Impulsen oder weißes Rauschen und das Filter besteht meistens aus einer Kette von Resonanzfiltern (Fant, 1970: 15ff).

Natürlich spielen die Lungen eine elementare Rolle, wie etwa auch die Blasebälge einer Orgel, ohne die kein Ton zustande käme. Die Lungen sind aber nur insofern von Bedeutung,

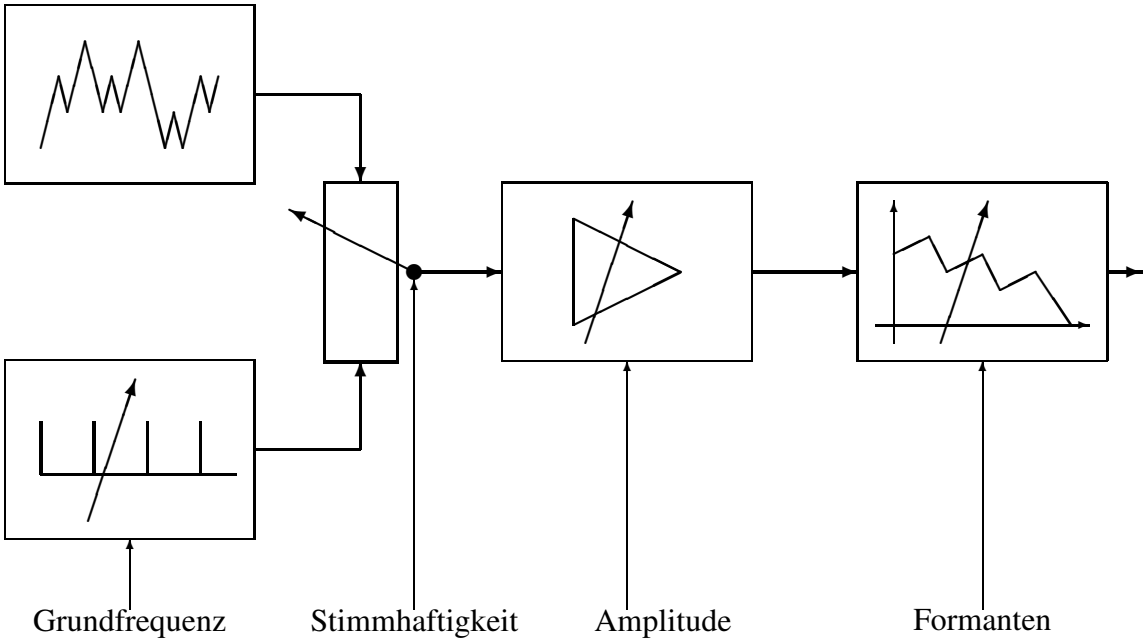

Abb. 10.2: Quelle-Filter-Modell der Sprachproduktion.

als daß sie durch den verursachten Luftstrom die Stimmbänder in Schwingung versetzen, vergleichbar mit dem Rohrblatt im Mundstück einer Klarinette, das auch bei strömender Luft durch den sog. Bernoulli-Effekt zu schwingen beginnt.

Der Aufbau und die Funktionsweise vieler Musikinstrumente ähnelt Teilen des physiologischen Aufbaus des menschlichen Sprechapparates. Trotzdem ist es mit keinem der bekannten Musikinstrumente möglich, eine verständliche Äußerung zu produzieren. Dies hat einen einfachen Grund: Kein Instrument ist in der Lage, seinen Klangcharakter derart schnell und zugleich gezielt zu verändern, wie dies beim Sprechen durch die Veränderungen des Ansatzrohres geschieht, die durch Lippen- und Zungenbewegungen hervorgerufen werden. Zudem verändert sich beim Sprechen noch die Tonhöhe des vom Kehlkopf erzeugten Stimmtons. Auch wird der Ton immer wieder durch Rauschen oder auch Stille unterbrochen, die für die Erzeugung von stimmlosen Konsonanten benötigt werden.

Um Sprache künstlich zu erzeugen, muß es also gelingen, die Schnelligkeit und gleichzeitige Gezieltheit klanglicher Veränderungen, durch die Sprache gekennzeichnet ist, zu simulieren. Bei den hier vorgestellten Versuchen wird die Variabilität der Tonhöhe durch einen konstanten oder langsam und gleichmäßig fallenden Ton ersetzt, da es für die Erkennung des Gesprochenen nur zweitrangig ist, ob ein Sprecher monoton oder mit bewegtem Stimmton spricht, denn Betonungen lassen sich auch durch Dauer und Intensität markieren.

Vielleicht ist zum Schluß dieser Vorüberlegungen noch der Vergleich mit einer Trompete, die mit einem Dämpfer gespielt wird, angebracht. Manchmal wird zur Symbolisierung von Sprache eine Trompete mit schnell auf und zu bewegtem Dämpfer verwendet, da die resultierende zeitliche Strukturierung des Klanges gewisse Ähnlichkeiten mit der der menschlichen Sprache hat, wenngleich kein Wort zu verstehen ist. Die Schnelligkeit der Klangveränderung wird zwar erreicht, die Gezieltheit jedoch keineswegs.

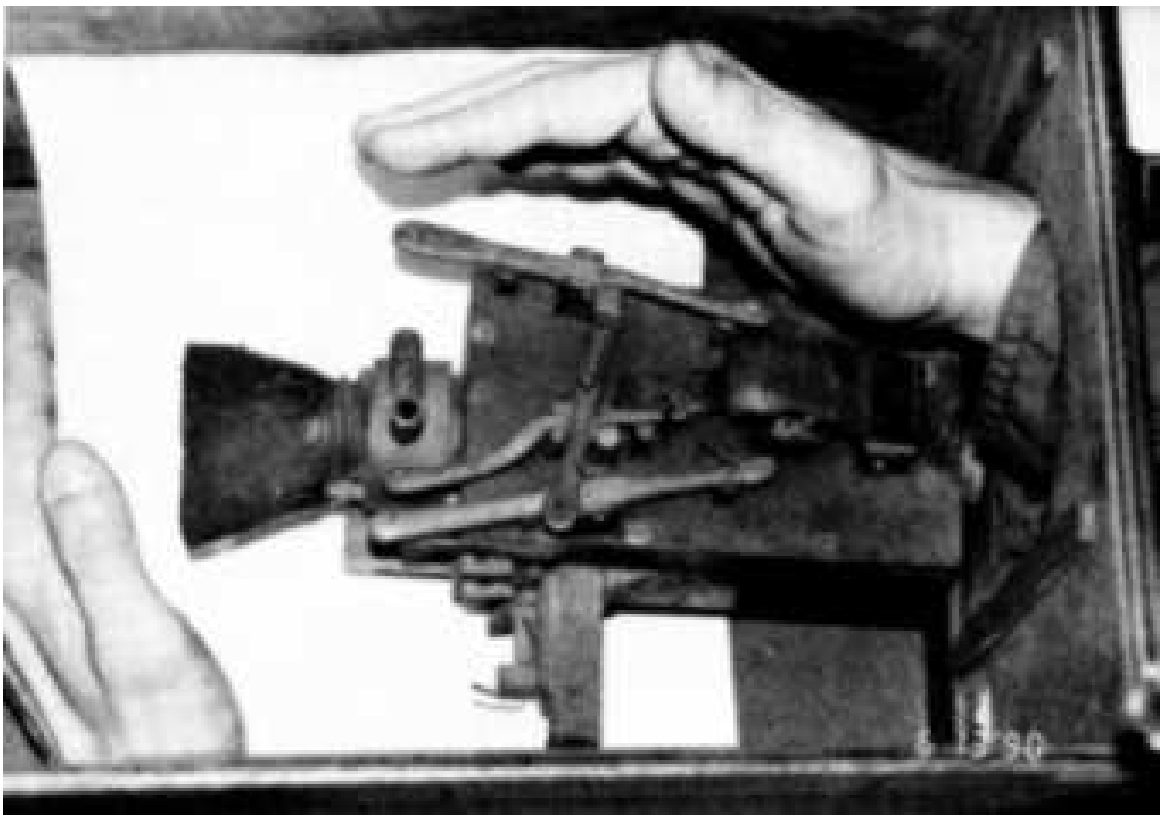

Abb. 10.3: Die Sprechmaschine von Wolfgang von Kempelen im Deutschen Museum München.

10.2 Die Sprechmaschine von Wolfgang von Kempelen

Wolfgang von Kempelen hat in der Zeit von etwa 1780 bis zur Veröffentlichung ihrer Beschreibung im Jahre 1791 eine Maschine gebaut (siehe Abb. 10.3 und Abb. 10.4), die bei richtiger Bedienung sprachliche Äußerungen und sogar ganze Sätze hervorbringen konnte. Allerdings war von Kempelen damals bereits bekannt und berüchtigt durch eine von ihm entwickelte Schachspielmaschine, die nicht etwa selbst Schach spielen konnte, wie man anfangs tief beeindruckt annahm, sondern die bedient werden mußte durch einen Schachspieler, der sich in ihr befand.[1] Demzufolge wurde von Kempelens Sprechmaschine in der damaligen Zeit sehr skeptisch begutachtet und rezensiert. Schließlich jedoch ließ sich jeder Kritiker davon überzeugen, daß diese Maschine wirklich sprechen konnte.

Von Kempelen begann seine Erforschung der sprechenden Maschine 1769 mit der Suche nach Musikinstrumenten, die dem Klang der menschlichen Stimme ähnlich waren.[2] Unter anderem untersuchte er auch das Waldhorn, die Trompete, die Maultrommel und verschiedene Mundstücke von Oboe, Klarinette und Fagott. Der Klang des ungarischen

1 Der erste Schachspieler in der Maschine stammte aus der Türkei, woraus sich dann die bis heute bekannten Redewendungen „etwas türken“ bzw. „einen Türken bauen“ entwickelt haben.

2 Von Kempelen schreibt, daß er mit diesen Nachforschungen bereits während der Arbeiten an dem Schachspieler begonnen hatte (1791: 390; Dudley & Tarnóczy, 1950). Damit lägen seine ersten Sprachsyntheseversuche zeitlich vor den Veröffentlichungen der Kratzensteinschen fünf „Vokalpfeifen“, die landläufig den Beginn der Sprachsyntheseforschung markieren (Kratzenstein, 1780 und 1782).

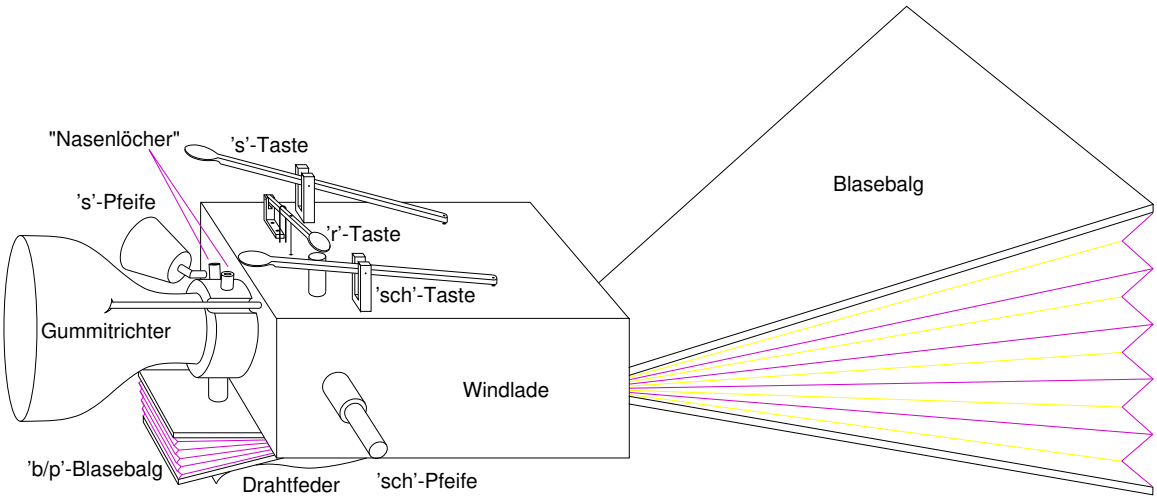

Abb. 10.4: Skizze der von Kempelenschen Sprechmaschine, angefertigt nach Originalkupfertafeln.

Dudelsacks kam seinen Vorstellungen am nächsten. Daher verwendete er dessen Rohrblatt als Schallquelle für die stimmhaften Laute. Nach langen Versuchsreihen mit unterschiedlich geformten, umschaltbaren Hohlräumen für die verschiedenen Vokale führte ein durch die Hand zu verformender Gummitrichter (Abb. 10.4) als Resonator zu brauchbaren Vokalqualitäten und -übergängen.

Die Kombination von Quelle und Filter zusammen mit weiteren Verfeinerungen zur Erzeugung von Konsonanten ergab schließlich eine Sprachqualität, die seinerzeit mit der Stimme eines „fünf- bis sechsjährigen Knaben“ (Windisch, 1783: 49) verglichen wurde, der ebenfalls gelegentlich unsauber artikuliert. Das System sprach vorwiegend in Französisch, „weil diese Sprache weniger Härten und zischende Laute hat als z.B. unsere Sprache“ (Hindenburg, 1784: 53). Statt der Plosivlaute „p“, „t“ und „k“ konnte das System nämlich prinzipiell nur einen „p“-ähnlichen Laut erzeugen. Auch die anderen Konsonanten machten dem System größere Schwierigkeiten als die Vokale. Hierauf werden wir später bei der Entwicklung unserer Synthesesteuerung noch einmal zurückkommen.

10.3 Problematik der Konkatenation von gesampleten Sprachteilen

Wolfgang von Kempelen erkannte bereits, daß Sprache nicht einfach eine Aneinanderreihung von einzelnen isoliert artikulierten Sprachlauten ist, sondern „daß um ganze zusammengesetzte Wörter herauszubringen, die Stimme nicht aus mehreren, sondern immer aus einer und derselben Röhre herausgehen müsse“ (Kempelen, 1791: 198f). Er folgerte, „daß ich schlechterdings der Natur folgen müßte, die nur *eine* Stimmritze, und nur *einen* Mund hat, zu dem alle Laute herausgehen, und eben nur darum sich miteinander verbinden“ (Kempelen, 1791: 407).

Die gegenwärtig moderne und sehr erfolgreiche Methode des Aneinanderhängens von gesampleten Sprachteilen (z.B. Phonen, Silben, Worten, u.s.w.) zur Erzeugung einer beliebigen sprachlichen Äußerung verstößt prinzipiell gegen diesen von Kempelenschen Grundgedanken. Das Konkatenieren von Sprachsamples ist wie das Nacheinandersprechen mit je-

weils unterschiedlich geformten und damit inkompatiblen Ansatzrohren. Die inadäquaten Übergänge zwischen den unterschiedlichen Konfigurationen führen zu nicht unerheblichen Verständnisproblemen, wenngleich die grundsätzliche Klangqualität dieser Sprachsynthesemethode in Akzeptanzuntersuchungen sehr hoch eingeschätzt wird. Diese Art der Synthese sprachlicher Äußerungen, die sich sicherlich mit einem Musik-Sampler verhältnismäßig einfach realisieren ließe, kommt für unsere Zwecke wegen der typisch fehlerhaften Laut- oder Wortübergänge nicht in Frage.

10.4 Akustische Analyse von Sprache durch den Spektrographen

Uns interessiert vielmehr, wie die Akustik der Sprache aussieht und welche akustischen Strukturen für das Verstehen von Sprache ausschlaggebend sind und folglich künstlich nachgebildet werden müssen.

Bereits seit Anfang des Zweiten Weltkrieges wurde an den *Bell Telephone Laboratories* unter der Leitung von Ralph Potter der *Sound Spectrograph* entwickelt und das Verfahren kurz nach Kriegsende 1946 im *Journal of the Acoustical Society of America* publiziert.[3] Das Gerät mißt die Energie in Abhängigkeit von Zeit sowie von Frequenz und stellt alle Meßpunkte eines Signals in einem Bild dar, dem sog. Spektrogramm oder auch Sonagramm. Die X-Achse des Sonagramms entspricht der Zeit und die Y-Achse der Frequenz. Die jeweilige Energie wird dabei in Graustufen kodiert: je dunkler ein Punkt in der Abbildung erscheint, desto größer ist die spektrale Kurzzeitenergie zu einem der X-Koordinate entsprechenden Zeitpunkt bei derjenigen Frequenz, die der Y-Koordinate entspricht (siehe z.B. Abb. 10.8).

Hierdurch war erstmals die Visualisierung der spektralen Eigenschaften von zeitveränderlichen Signalen möglich. Gerade die Akustik von Sprache, deren Klangcharakter und damit spektrale Zusammensetzung sich von Augenblick zu Augenblick verändert, ließ sich nun endlich einer präzisen Analyse unterziehen. Aber auch nach zahlreichen Untersuchungen mit diesem Gerät blieb letztendlich die Frage offen, welche akustischen Eigenschaften für die Wahrnehmung und Identifizierung von Lauten relevant waren. Es zeigte sich nämlich, daß wörtlich gleiche Äußerungen von verschiedenen Sprechern zu ebenfalls sehr unterschiedlichen Abbildungen führen konnten.

10.5 Pattern-Playback-Synthese

Erst durch die Erfindung und den Bau eines inversen *Sound Spectrographen*, der sog. *Pattern-Playback-Maschine*, konnten auf opto-elektrischem Wege Sonagramme — und eben auch gezielt retuschierte Sonagramme — und deren stilisierte Zeichnungen mit akzeptabler Klangqualität wieder hörbar gemacht werden.

Diese Maschine[4] (siehe Abb. 10.5) wurde etwa 1950 von Franklin Cooper und Alvin Liberman an den *Haskins Laboratories* in New York (heute in New Haven) gebaut und in

3 Stellvertretend für eine ganze Reihe von Publikationen seien hier die drei wichtigsten genannt: Potter, 1946; Koenig, Dunn und Lacey, 1946; Potter, Kopp und Green, 1947.

4 Auch hier muß wohl Ralph Potter als der Erfinder angesehen werden, da bereits 1947 ein Patent zum Funktionsprinzip auf seinen Namen registriert war (U.S. Patent No. 2,432,123; Cooper et al., 1951: 325).

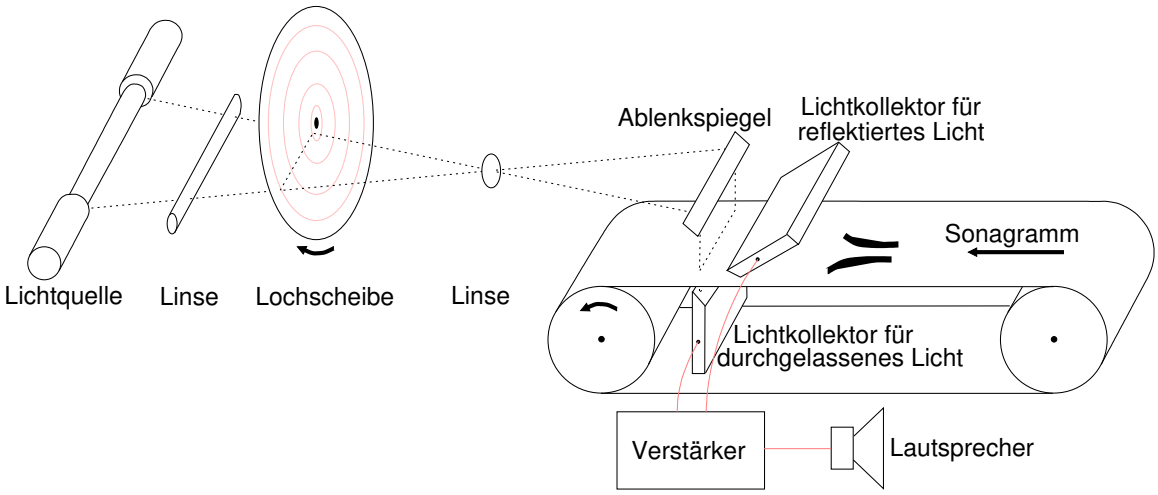

Abb. 10.5: Eine Skizze des Funktionsprinzips der Pattern-Playback-Maschine.

vielen bedeutenden Wahrnehmungsexperimenten der fünfziger und sechziger Jahre angewendet, die das phonetische Wissen über die Beziehungen zwischen Akustik und Wahrnehmung entscheidend erweitert haben.[5]

In Abb. 10.6 sehen wir das Negativ eines Originalsonagramms sowie eine handgezeichnete Variante desselben, die sich auf die wesentlichen akustischen Merkmale beschränkt, welche für das Verstehen der Äußerung notwendig sind. Da die *Pattern-Playback-Maschine* mit einer konstanten Stimmtonfrequenz von 120 Hz arbeitete, wurden durch die Resynthese die Bewegungen des Stimmtons der ursprünglichen Äußerung eliminiert. Weitere Details zum Funktionieren der Maschine finden sich bei Flanagan (1972: 213f).

An dieser Stelle sollte festgehalten werden, daß diese Maschine ein Synthesesystem darstellt, das einem Zeichner mit ausreichendem phonetischem Wissen über die Bedeutung akustischer Merkmale erlaubt, beliebige sprachliche Äußerungen zu generieren.

10.6 Neuere Sprachsyntheseverfahren

Bevor wir uns nun selber als „Zeichner“ von Sprachklängen betätigen, wollen wir einen abschließenden Blick auf die weitere historische Entwicklung der Syntheseverfahren werfen.

Mit der *Pattern-Playback-Maschine* endete die Ära der mechanischen Synthesesysteme Ende der sechziger Jahre. Doch bereits in den dreißiger Jahren begann eine neue Entwicklung, die der elektrischen Systeme. Diese wurden anfangs mit Röhren und später, nach Entdeckung der Halbleiter, mit Transistoren aufgebaut. Gegenwärtig werden die Systeme nur noch in Software für handelsübliche Computer oder Spezial-Hardware realisiert.

Die ersten bedeutenden Sprachsynthesegeräte waren der *Vocoder*[6], der 1936 an den bereits erwähnten *Bell Telephone Laboratories* von Homer Dudley entwickelt wurde, und der

5 Dem interessierten Leser sei hierzu das umfassende Werk von Liberman (1996) empfohlen.

6 *Vocoder* bedeutet voice coder (Dudley, 1936; Meyer-Eppler, 1949: 116ff).

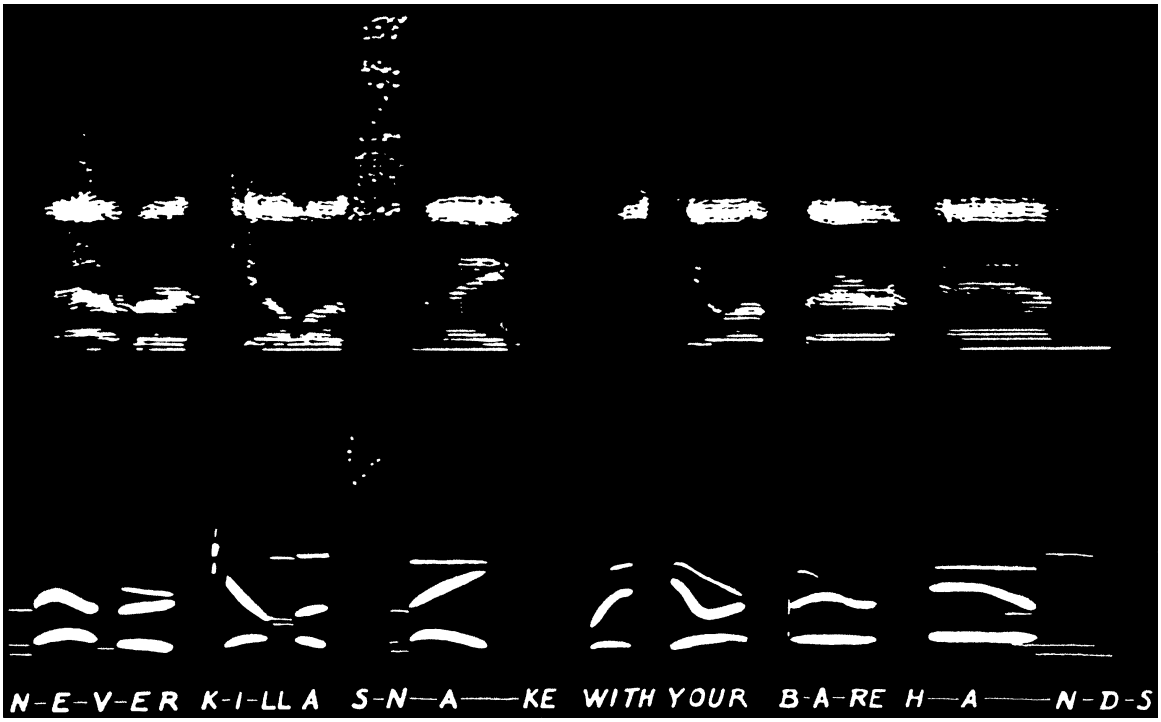

Abb. 10.6: Negativ eines Sonagramms (oben) und zugehörige handgezeichnete Stilisierung (unten).

Voder[7], der erstmals 1939 öffentlich vorgestellt wurde und sich allein durch eine Tastatur steuern ließ. Beide Geräte modellierten das Frequenzspektrum durch etwa zehn Bandpaßfilter. Die in heutigen Musik-Synthesizern wiederentdeckten Vocoder verwenden ebenfalls Filterbänke, allerdings mit bis zu etwa 32 Bandpaßfiltern.

1970 begann Dennis Klatt mit der Entwicklung eines Software-Synthesizers, der auf dem Formantsynthesizer von Lawrence aus dem Jahre 1953 basierte, sich aber durch Parameterdateien steuern ließ. 1979 entwickelte sich dieses erfolgreiche Konzept zum *MITalk*-Synthesizer, 1981 zum *Klattalk*-Synthesizer und schließlich 1983 zum *Digital DECtalk*.[8]

Je mehr sich bei der Sprachsynthese das Konzept der Formanten durchsetzte, die z.B. in Abb. 10.6 gerade bei der Handzeichnung deutlich hervortreten, desto einfacher und zugleich präziser wurden die Synthesesysteme, da sie für jeden zu modellierenden Formanten nur einen variablen Bandpaßfilter benötigten. Vokalische Klangverläufe, die z.B. bei Diphthongen wie /au/ oder /ai/, aber auch bei vokalisierten Konsonanten auftreten (z.B. wird das Wort *Eier* nicht [air], sondern [aia] ausgesprochen), lassen sich für gute Sprachverständlichkeit durch nur zwei Formanten realisieren, deren spektrale Verläufe in der Zeit allerdings sehr genau eingestellt werden müssen.

Am Rande sei erwähnt, daß die Sprachübertragung mit stark reduzierten und komprimierten Datenmengen, wie sie u.a. bei Mobiltelefonen verwendet wird, nicht als Sprachsynthese bezeichnet werden kann, da die Datenströme keinerlei sinnvolle Manipulation zulassen, um gezielt sprachliche Äußerungen hervorzubringen.

---

7 Dudley entwickelte auch den *Voder* (voice operation demonstrator, Dudley, Riesz und Watkins, 1939).

8 Das Funktionsprinzip und die geschichtliche Entwicklung sind ausführlich in Klatt (1987) dargestellt.

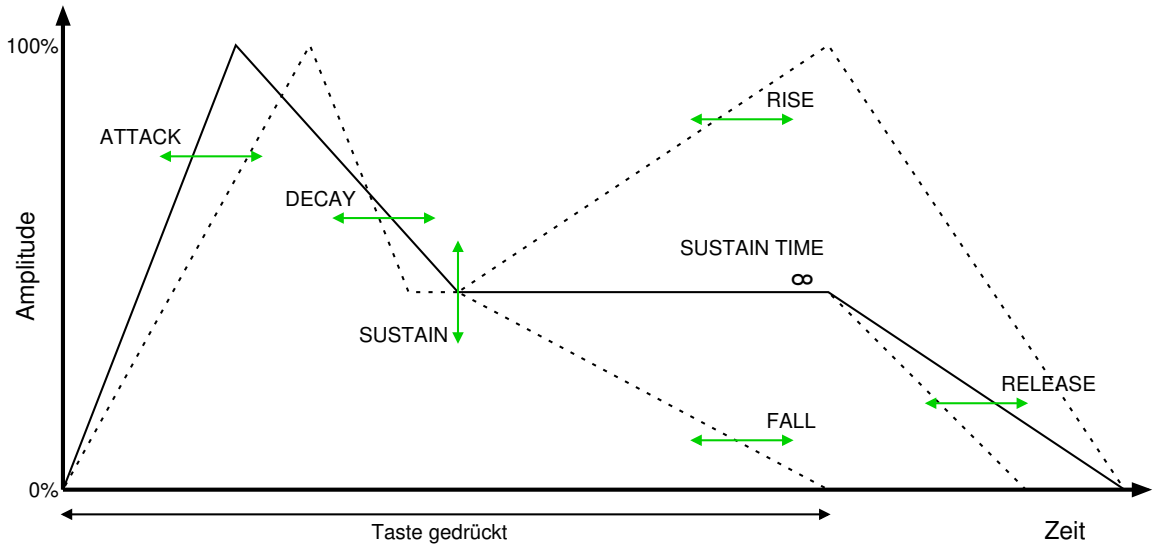

Abb. 10.7: Einstellungsmöglichkeiten des Hüllkurvenverlaufs beim *Access Virus*.

10.7 Subtraktive Synthese — Möglichkeiten des Analog-Synthesizers

Es muß einmal klar gesagt werden, daß das Quelle-Filter-Modell der Phonetik ein Spezialfall der subtraktiven Synthese ist, die die analogen Musik-Synthesizer geradezu charakterisiert. Ein obertonreiches Signal wird durch ein Filter in seinem spektralen Verlauf beeinflußt; dabei kann das Filter keine Obertöne hinzufügen, die nicht schon im Quellsignal vorhanden sind, sondern es kann lediglich Frequenzbereiche in ihrer Amplitude anheben (und auf diese Weise auch Formanten erzeugen) oder absenken bzw. gänzlich herausfiltern.

In einem typischen analogen Synthesizer wird die Quelle realisiert, indem ein bis drei Oszillatoren mit wählbaren Schwingungsformen („Sägezahn“, „Rechteck“, „Dreieck“, „Weißes Rauschen“, u.s.w.) und Frequenzverhältnissen gemischt und einem oder zwei Filtern mit 24 dB Flankensteilheit und Resonanzfähigkeit zugeführt werden. Diese Konstellation sollte aber unter Berücksichtigung des bisher dargestellten Wissens über die akustische Struktur von Sprache hinreichend sein, um sprachliche Äußerungen zu generieren, denn wir wissen ja bereits, daß zur Synthese von verständlichen vokalischen Äußerungen die Kontrolle von zwei Formantfrequenzen benötigt wird.

Problematisch ist vielmehr die adäquate Steuerung der Filterfrequenzverläufe in der Zeit, da die meisten Musik-Synthesizer nur verhältnismäßig eingeschränkte Möglichkeiten haben, zeitliche Frequenz- oder Amplitudenveränderungen zu erzeugen. Zu diesem Zweck verfügen sie über sog. Hüllkurvengeneratoren, die zeitlich variable Modulationsdaten erzeugen können (siehe Abb. 10.7). Oft stehen hier nur vier bis sechs Kontrollparameter (Attack, Decay, Sustain, Release, u.a.)[9] zur Verfügung, die zwar für die Imitation von Zeitverläufen bei natürlichen Musikinstrumenten vollkommen ausreichen, für komplexe sprachliche Äußerungen jedoch nicht hinreichend sind (siehe Formantverläufe in Abb. 10.6). Wir werden uns also darauf beschränken, exemplarisch einen Diphthong (das Wort *Ei*) und einen Triphthong (das Wort *Eier*) zu programmieren.

9 Hüllkurvengeneratoren werden oft auch, die Parameternamen abkürzend, ADSR-Generatoren genannt.

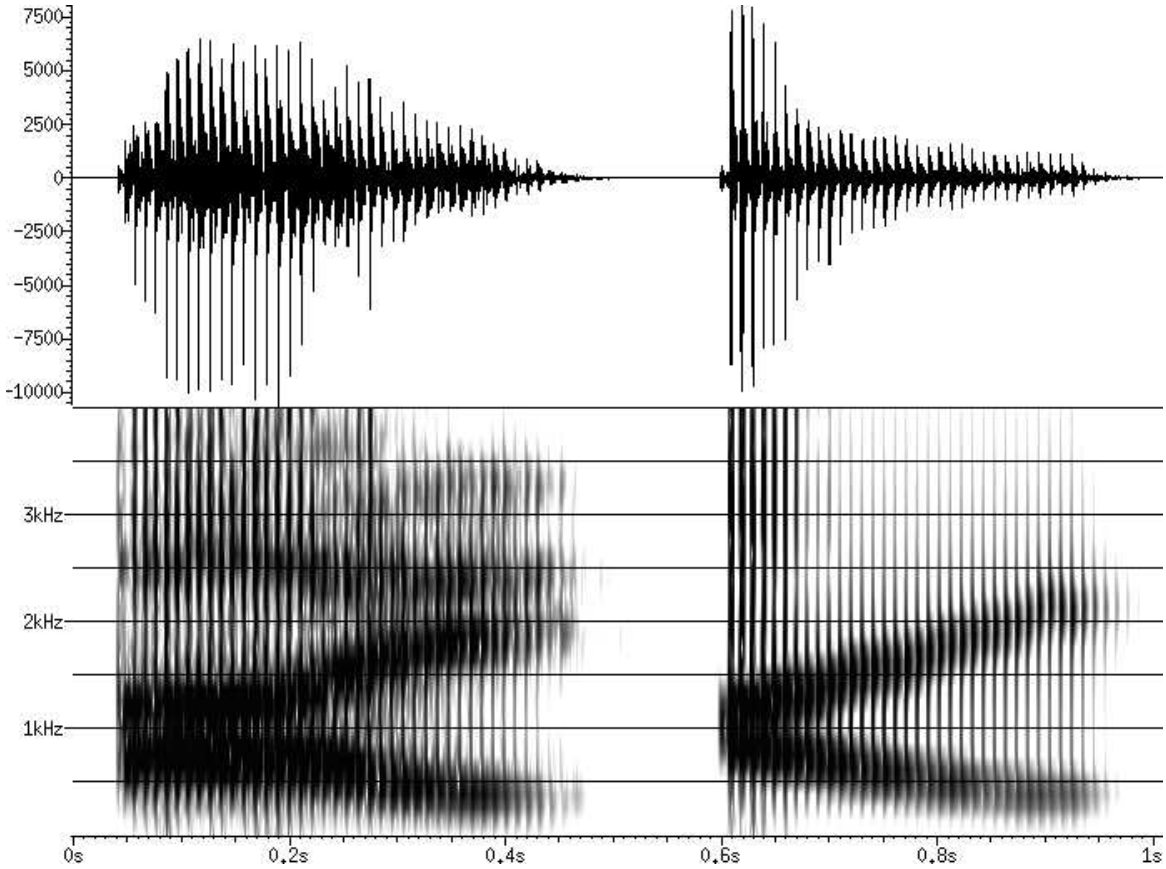

Abb. 10.8: Oszillogramme und Sonagramme des Diphthongs /ai/, gesprochen von einem Menschen (links) und erzeugt mit einem Musik-Synthesizer (rechts).

10.8 Synthese von Vokalen und Diphthongen

Wie schon von Kempelen können auch wir Konsonanten nur mit größtem Aufwand hervorbringen, da sie im Vergleich zu Vokalen komplexere zeitliche und/oder spektrale akustische Strukturen erfordern, die mit den eben beschriebenen beschränkten Möglichkeiten eines Musik-Synthesizers nur mangelhaft realisierbar sind. Insbesondere die Veränderungen des Quellsignals etwa bei Plosivlauten, aber auch die Änderungen des Filters bei Übergängen etwa zwischen Vokalen und Nasallauten sind sehr komplex. Wir beschränken uns also zunächst auf reine Vokale und Diphthonge.

In Abb. 10.8 ist die spektrographische Analyse des Wortes *Ei* zu sehen. Links wurde eine menschliche Äußerung analysiert: Sehr gut sind die ersten beiden Formanten zu erkennen, auf die wir uns hier beschränken müssen, da der verwendete Synthesizer *nur* über zwei Filter verfügt. Während der erste Formant bei etwa 700 Hz einsetzt, um dann bis auf etwa 350 Hz abzusinken, startet der zweite Formant bei 1.2 kHz und steigt dann kontinuierlich auf 1.9 kHz an.

Alle Hüllkurven- und die Filterparameter *Cutoff-Frequency, Resonance* und *Modulationsintensität* des Synthesizers müssen nun durch *trial-and-error* so eingestellt werden, daß das spektrographische Analyseergebnis etwa mit dem Vorbild übereinstimmt. Auch sollte

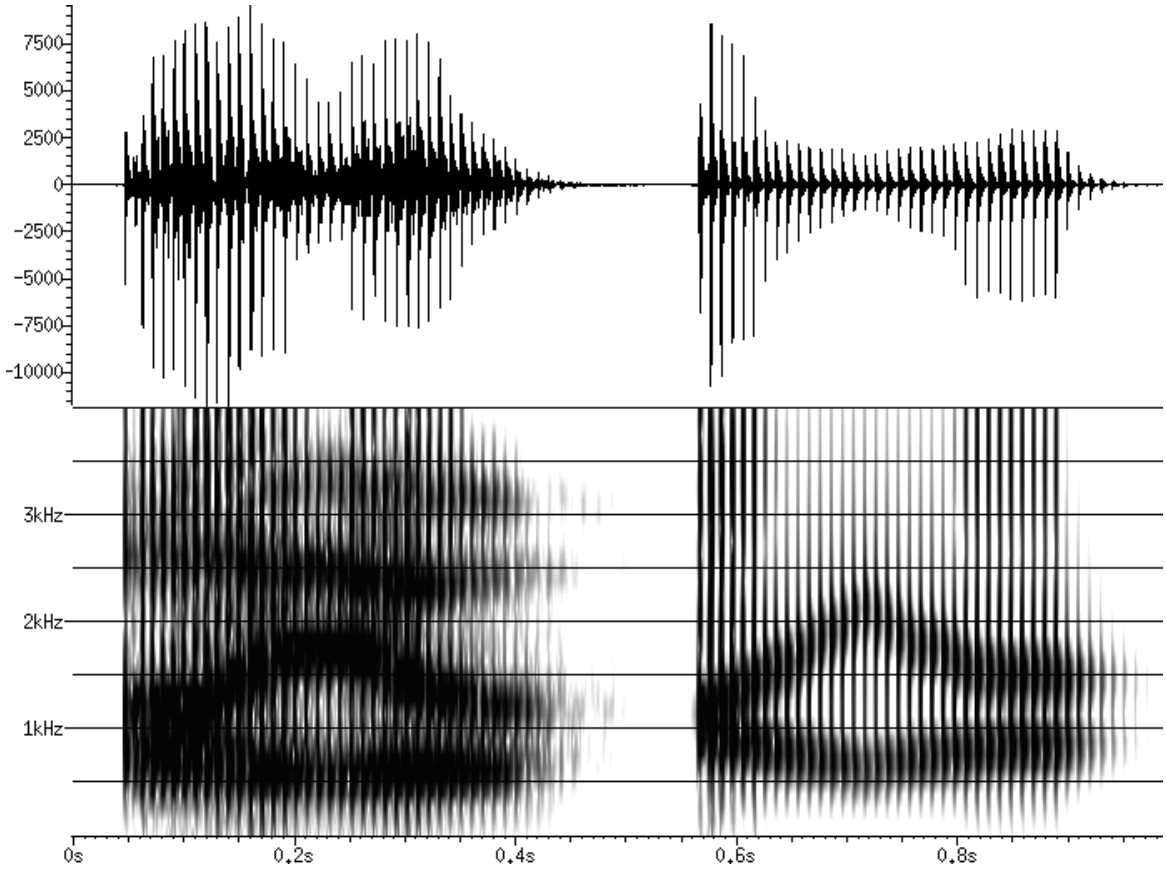

Abb. 10.9: Oszillogramme und Sonagramme des Wortes *Eier*, gesprochen von einem Menschen (links) und erzeugt mit einem Musik-Synthesizer (rechts).

das synthetische Signal immer wieder akustisch kontrolliert werden, denn z.B. in unserem Fall zeigt sich, daß die Endfrequenz des zweiten Formanten besser bei 2.2 kHz gewählt werden sollte, da der im originalen Sprachsignal deutlich zu erkennende dritte Formant mit dem zweiten Formanten am Ende des Diphthongs synergetisch zusammenwirkt und dadurch die Wahrnehmung einer höheren als der physikalischen Frequenz hervorruft.

10.9 Synthese eines Wortes

Wir erweitern nun den eben generierten Diphthong /ai/ zu dem Triphthong /aia/, der dem Wort *Eier* entspricht. Auch hier sind bei der Adjustierung der Klangparameter wieder einige Besonderheiten der akustischen Struktur von Sprache zu beachten. So fällt in Abb. 10.9 zunächst auf, daß der erste Formant hier nicht mehr bis auf 350 Hz abfällt, sondern nur etwa bis auf 550 Hz. Auch entsprechen die Endpositionen der beiden Formanten nicht den Anfangspositionen, obwohl der erste und der letzte Laut des Wortes sehr große phonetische Ähnlichkeit besitzen.

Wenn ein Mensch Laute zu Worten verbindet und artikuliert, dann weichen die akustischen Eigenschaften der isolierten Laute mehr oder weniger von ihren Eigenschaften in

verbundener Sprache ab. Dieser Effekt nennt sich *Koartikulation* und führt zu einer gewissen *Enkodiertheit* der akustischen Merkmale: sie verlieren an visueller Offensichtlichkeit, indem sie sich auch in den benachbarten Lauten manifestieren. Für den Hörer resultieren daraus keine Verständnisprobleme; das Gegenteil ist sogar der Fall: Dadurch, daß sich Laute bereits in ihren Nachbarlauten ankündigen, erhält das Gehör mehr Zeit zur Extraktion der Merkmale und zur Erkennung einzelner Laute als bei rein sequenzieller Darbietung und kann Sprache dadurch besser dekodieren.

Für ein akzeptables akustisches Ergebnis ist es also notwendig, diese Phänomene zu kennen und zu modellieren. Eine einfache Konkatenation von Sprachsamples würde an dieser Stelle gravierende Fehler verursachen. Obwohl sich größere visuelle Unterschiede zwischen dem Original und dem synthetisierten Signal beobachten lassen, die insbesondere durch das Fehlen des dritten und vierten Formanten, aber auch durch einen nicht ganz stimmigen Amplitudenverlauf verursacht werden, ruft die Akustik dieser Signale doch einen bestechend menschlichen Eindruck hervor.

Natürlich ist unverkennbar, daß es sich um eine synthetisierte Äußerung handelt. Als Signalquelle wurde eine Sägezahnschwingung mit konstanter Frequenz (etwa 105 Hz) verwendet, während die Schwingungen der Stimmbänder eines Menschen weder sägezahnförmig noch exakt periodisch verlaufen. Auch die Modulation der Tonhöhe des Oszillators durch Rauschen würde nicht zu der Quasi-Periodizität führen, die für die menschliche Sprache so typisch ist. Wir könnten unser Modell an dieser Stelle verfeinern, doch sind bereits hier klar die Begrenzungen in Sicht, denen ein Musik-Synthesizer bei der Sprachsynthese unterworfen ist.

10.10 Abschließende Bemerkungen

Die hier vorgestellten Versuche, mit einem handelsüblichen Musik-Synthesizer durch gezielte Programmierung sprachliche Äußerungen hervorzubringen, sollten vor allem als Denkanstoß und Inspiration zu eigenen Klangexperimenten verstanden werden. Durch leichte Variationen verschiedener Hüllkurven- oder Filterparameter läßt sich eine Vielzahl von interessanten Klängen kreieren. Eine ganz andere Möglichkeit, der Beschränkung durch die Hüllkurvengeneratoren zu entgehen, besteht in der Verwendung von MIDI-Controller-Daten, die von einem Sequenzer geliefert oder sogar in Echtzeit aus einem Sprachsignal extrahiert werden könnten. Die daraus resultierende Klangfamilie wird an anderer Stelle beschrieben werden.

Die von mir hier vorgestellten Klänge liegen auf meiner Homepage[10] im Standard MIDI-File-Format vor. Sie können mit einer beliebigen Sequenzer-Software in den *Access Virus* geladen werden.

Der historisch interessierte Leser sei auf die weiterführenden Werke von Meyer-Eppler (1949), Flanagan (1972), Flanagan & Rabiner (1973), Klatt (1987), Pompino-Marschall (1991) und Tillmann (1994) verwiesen, während der aktuelle Stand der Sprachsyntheseforschung anhand der Publikationen von Bailly & Benoît (1992), Santen et al. (1997), Dutoit (1997) und Sproat (1998) erfaßt werden kann.

---

10 `http://www.phonetik.uni-muenchen.de/~hpt/kf/`

Literatur

Bailly, G.; Benoît, C. (Hrsg.). (1992). *Talking Machines — Theories, Models, and Designs*. North-Holland, Amsterdam.

Cooper, F. S.; Liberman, A. M.; Borst, J. M. (1951). The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proc. of the National Academy of Sciences*, 37(5): S. 318–325.

Dudley, H. (1936). Synthesizing speech. *Bell Laboratories Record*, 15: S. 98–102.

Dudley, H.; Riesz, R. R.; Watkins, S. S. A. (1939). A synthetic speaker. *J. of the Franklin Institute*, 227: S. 739–764.

Dudley, H.; Tarnóczy, T. H. (1950). The speaking machine of Wolfgang von Kempelen. *J. of the Acoustical Society of America*, 22(2): S. 151–166.

Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht.

Fant, G. (1970). *Acoustic Theory of Speech Production*. Mouton & Co., The Hague; Niederlande, 2. Aufl. (1. Aufl.: 1960).

Flanagan, J. L. (1972). *Speech analysis, synthesis and perception*, Band 3 aus *Kommunikation und Kybernetik in Einzeldarstellungen*. Springer-Verlag, Berlin, 2. Aufl. (1. Aufl.: 1965).

Flanagan, J. L.; Rabiner, L. R. (Hrsg.). (1973). *Speech Synthesis*. Dowden, Hutchinson & Ross, Stroudsburg, Pennsylvania.

Hindenburg, C. F. (1784). *Ueber den Schachspieler des Herrn von Kempelen. Nebst einer Abbildung und Beschreibung seiner Sprachmaschine*. Mathematische Fakultät der Universität, Leipzig.

Kempelen, W. R. v. (1791). *Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine*. J. V. Degen, Wien.

Klatt, D. H. (1987). Review of text-to-speech conversion for english. *J. of the Acoustical Society of America*, 82(3): S. 737–793.

Koenig, W.; Dunn, H. K.; Lacey, L. Y. (1946). The sound spectrograph. *J. of the Acoustical Society of America*, 18(1): S. 19–49.

Kratzenstein, C. G. (1780). *Tentamen coronatum de voce*. Acta Acad., Petrograd.

Kratzenstein, C. G. (1782). Sur la naissance de la formation des voyelles. *J. de Physique*, 21: S. 358–380.

Liberman, A. M. (1996). *Speech: A Special Code*. MIT Press, Cambridge, Massachusetts.

Meyer-Eppler, W. (1949). *Elektrische Klangerzeugung. Elektronische Musik und synthetische Sprache*. Ferd. Dümmlers Verlag, Bonn.

Pompino-Marschall, B. (1991). Wolfgang von Kempelen und seine Sprechmaschine. Eine biographische Notiz zum 200. Jahrestag der Publikation seines „Mechanismus der mensch-

lichen Sprache“. Forschungsberichte (FIPKM) 29, Institut für Phonetik und Sprachliche Kommunikation der Universität München, S. 181–252.

Potter, R. K. (1946). Introduction to technical discussions of sound portrayal. *J. of the Acoustical Society of America*, 18(1): S. 1–3.

Potter, R. K.; Kopp, G. A.; Green, H. C. (1947). *Visible Speech*. D. van Nostrand, New York.

Santen, J. P. H. v.; Sproat, R. W.; Olive, J. P.; Hirschberg, J. (Hrsg.). (1997). *Progress in Speech Synthesis*. Springer-Verlag, New York.

Sproat, R. W. (Hrsg.). (1998). *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Kluwer Academic Publishers, Dordrecht.

Tillmann, H. G. (1994). Phonetics, early modern, especially instrumental and experimental work. In: Asher, R. E.; Simpson, J. M. Y. (Hrsg.), *The Encyclopedia of Language and Linguistics*, Band 6, S. 3082–3095. Pergamon Press, Oxford.

Windisch, K. G. v. (1783). *Karl Gottlieb von Windisch's Briefe über den Schachspieler des Herrn von Kempelen nebst drey Kupferstichen, die diese berühmte Maschine vorstellen*. Mechel, Christian von, Basel.