HEINRICH-HERTZ-INSTITUT FÜR SCHWINGUNGSFORSCHUNG
BERLIN-CHARLOTTENBURG

Technischer Bericht Nr. 137

Rechengesteuerte Spracherzeugung

von

Claus-Eberhard Liedtke

Berlin
1971

Technischer Bericht Nr. 137

Rechnergesteuerte Spracherzeugung

Zusammenfassung:

Für einen Digitalrechner mittlerer Größe soll eine Sprachausgabe entwickelt werden. Um die Sprache ökonomisch abspeichern zu können, muß man sie vorher komprimieren, d.h. von ihrer Redundanz befreien. Es werden zunächst verschiedene Möglichkeiten der Sprachkompression beschrieben und miteinander verglichen. Dabei wird der Formantvocoder als das Prinzip ermittelt, das die günstigsten Voraussetzungen im Zusammenhang mit der vorliegenden Aufgabenstellung bietet. Sowohl die Simulation des Formantvocoders auf einem Digitalrechner als auch Entwurf zur hardwaremäßigen Realisierung des Synthetisators werden ausführlich beschrieben.

H e i n r i c h - H e r t z - I n s t i t u t f ü r S c h w i n g u n g s f o r s c h u n g

Der Bearbeiter

C.-E. Liedtke

(Liedtke)

Der Abteilungsleiter

(Prof. Dr. Ing. W. Giloi)

Der Institutsdirektor

(Prof. Dr. Ing. Gundlach)

Berlin-Charlottenburg, den 17. Dezember 1971

RECHNERGESTEUERTE SPRACHERZEUGUNG

von

Dipl.-Ing. Claus-Eberhard Liedtke

Diese Schrift wurde dem Fachbereich Elektrotechnik (FB 19)
der Technischen Universitaet Berlin
am 14.7.1971 zur Annahme als Dissertation vorgelegt

I N H A L T S A N G A B E

1. Aufgabenstellung und Ergebnisse

Digitalrechner finden heute in allen Zweigen der Wirtschaft einen immer groesseren Einsatz. In absehbarer Zeit wird fast jeder Mensch in der einen oder anderen Weise mit einem Rechner zusammenarbeiten muessen, sei es im Zusammenhang mit der automatischen Lohnabrechnung, im vollautomatischen Bankverkehr, der rechnergesteuerten Boersenkursdurchgabe oder Wettervorhersage, der vollautomatischen Reisevermittlung, oder durch ein computergesteuertes Lehrgeraet oder Informationssystem. Die Reihe der Anwendungsbeispiele aus dem taeglichen Leben, in denen in naher Zukunft altgewohnte Organisationsformen von Rechnern abgeloest werden, liesse sich beliebig fortsetzen.

Die Einfuehrung von Computern in diesem Umfang ist aber nur dann moeglich, wenn diese benutzerfreundlich gemacht werden koennen. Der Grosseinsatz von Rechnern kann nicht erfolgen, wenn seine Bedienung nur durch eine kleine Anzahl ausgebildeter Spezialisten, den Programmierern und Operateuren, moeglich ist oder Zahlenkolonnen als Ausgabewerte des Rechners nur von Ingenieuren gedeutet werden koennen.

Aus diesem Grunde ist es noetig, dass der Rechner ausser gedruckter Information auch graphische und akustische Information verarbeiten und ausgeben kann. Eine Teilaufgabe der Rechnerentwicklung muss es daher, um der verbesserungsbeduerftigen Mensch- Maschine- Beziehung willen sein, dem Rechner die Sprache des Menschen beizubringen.

Eine weitere Ursache fuer die Forderung nach der Entwicklung einer Sprachausgabe fuer einen Digitalrechner ist oekonomischer Art. Man denke beispielsweise an die Errichtung eines umfangreichen Informationssystems. Dieses wird von einem zentralen Rechner gesteuert. Mit den technischen Moeglichkeiten, die heute gaengig sind, muesste jeder Benutzer dieses Informationssystems ueber ein Terminal verfuegen, das beispielsweise aus einer Fernschreibmaschine bestehen kann. Die Fernschreibmaschine wuerde durch ein Modem oder einen Telephon- Coupler ueber das Telephonnetz an den zentralen Rechner des Informationssystems angeschlossen sein. Da ein derartiges Terminal, auch bei groesserer Stueckzahl, kaum unter 4000.-- DM kostet, wird der Benutzerkreis des Informationssystems automatisch eingeschraenkt.

Wenn es jedoch genuegt, die Information akustisch weiterzugeben, so kann man ohne zusaetzlichen Kostenaufwand alle diejenigen in den Benutzerkreis aufnehmen, die ueber ein handelsuebliches Telephon verfuegen.

Der Verfasser hat es sich zur Aufgabe gemacht, aus den genannten Gruenden eine Sprachausgabe fuer einen Rechner zu entwickeln. Es soll sich dabei um eine Sprachausgabe fuer einen mittleren Digitalrechner handeln, bei der lediglich ein Standardtext ausgegeben werden soll. Das waere z.B. bei der oben genannten automatischen Boersenkursdurchsage oder einer rechnergesteuerten Lagerbestandueberpruefung der Fall.

Die Gesamtdauer des verfuegbaren Textes moege etwa 200 sek umfassen, das System soll aber auch fuer laengere Texte leicht ausbaufaehig sein.

Bei den oben genannten Anwendungsbeispielen kommt es nicht darauf an, dass die Sprache so klingt, als ob sie von einem Menschen gesprochen wuerde. Sie kann durchaus 'roboterhaft' klingen, wenn sie nur ausreichend verstaendlich ist.

Man koennte sich beispielweise vorstellen, dass eine derartige Sprachausgabe aus einem einfachen Tonband besteht, bei dem die Anfaenge der einzelnen auszugebenden Worte auf einer zweiten Spur gesondert markiert sind. Der Nachteil einer derartigen seriellen Sprachspeicherung besteht darin, dass es unter Umstaenden etliche Sekunden dauert, bis das Tonband bis zu der gewuenschten Stelle vor- oder zurueckgespult worden ist. Ein Benutzer eines derartigen Informationssystems erwartet aber sofort eine Antwort und wuerde bereits nach wenigen Sekunden Wartezeit den Hoerer auflegen in der Annahme, dass das System defekt sei.

Diese Schwierigkeit der seriellen Speicherung gesprochenen Textes laesst sich dadurch umgehen, indem man auf eine Trommel von ca 0.5 sek Drehzeit entsprechend viele parallele Tonspuren aufbringt. Die laengste Zugriffszeit fuer ein einzelnes Wort betraegt dann 0.5 sek. Bei 200 sek Sprache muessten auf der Trommel 400 parallele Spuren aufgebracht werden. Die Sprachausgabe liesse sich dadurch erweitern, dass entsprechend viele Trommeln parallel betrieben wuerden. Ein solches Spracherzeugungssystem ist sehr aufwendig und nur bei einem sehr kleinen Wortschatz wirtschaftlich.

Eine beliebig kleine Zugriffszeit fuer die einzelnen Worte laesst sich bei einem Rechner nur dann erreichen, wenn der gesprochene Text im Kernspeicher des Rechners gespeichert vorliegt. Dazu muss die gesprochene Sprache vorher quantisiert werden.

Wenn man gesprochene Sprache mit beispielsweise 10 kHz abtastet und die Amplitudenwerte mit 8 bit quantisiert, braucht man zur Speicherung von 1 sek Sprache 80000 bit. Bei einem Plattenspeicher mit einer mittleren Zugriffszeit von 40 ms und einer Lesegeschwindigkeit von 1 Million bit/sek wuerde es ca 120 ms dauern, bis ein Wort von 1 sek Dauer ausgegeben werden koennte.

Wenn ein derartiger Plattenspeicher voll fuer die Zwekke einer Sprachausgabe belegt wuerde, koennten bei einer Speicherkapazitaet von 24 Mbit auf diese Weise 300 s Sprache abgespeichert werden.

Wie Experimente gezeigt haben, ist der Mensch nicht in der Lage, Informationsraten von mehr als 50 bit/sek zu verarbeiten (/2/ S.7). Demzufolge muss der Informationsgehalt gesprochener Sprache kleiner als 50 bit/sek sein. Aus dem Verhaeltnis des oben genannten Speicherbedarfs der Sprache von 80000 bit/sek und des Informationsgehaltes von weniger als 50 bit/sek ergibt sich, dass die Sprache stark redundant

ist. Wenn es gelingt, die menschliche Sprache zu komprimieren, d.h. sie von ihrer Redundanz zu befreien, wird eine wesentlich oekonomischere Speicherung moeglich.

Sprachkompressionssysteme sind seit laengerer Zeit unter dem Namen Vocoder bekannt. Es wurden vom Verfasser in dem Zusammenhang verschiedene Vocoderarten auf ihre Brauchbarkeit fuer den vorliegenden Anwendungszweck untersucht. Die verschiedenen Vocoderarten unterscheiden sich hinsichtlich des Sprachkompressionsfaktors, der Wirtschaftlichkeit in Bezug auf die benoetigte Rechenzeit und dem Aufwand, der zum Aufbau einer in Realzeit arbeitenden Sprachausgabe als Hardwareausfuehrung notwendig ist.

Der Formantvocoder erschien dem Verfasser als das Vocodersystem, das am besten den gestellten Anforderungen gerecht wird und ausserdem die besten Voraussetzungen zur Erweiterung der Sprachausgabe fuer einen sehr grossen Wortschatz bietet.

Im Hinblick auf eine spaetere hardwaremaessige Realisierung des Syntheseteils wurde ein in seinem Aufbau besonders einfacher Formantvocoder entwickelt.

Die zur Steuerung der Synthesevorrichtung benoetigte Informationsrate betraegt je nach Sprachqualitaet 500 bis 2000 bit/sek.

Abb.1 zeigt in Form einer Visible- Speech- Darstellung das Wort 'HAWAII' bei verschiedenen Uebertragungsraten.

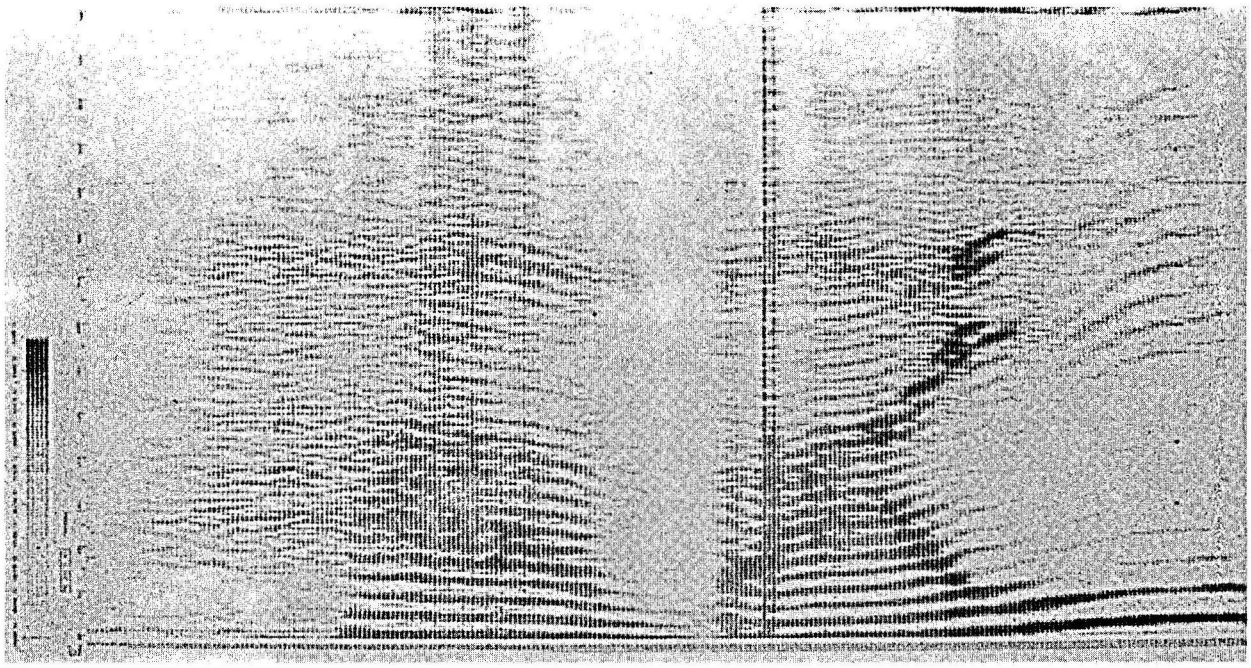


Abb.1a, 'HAWAII', Original

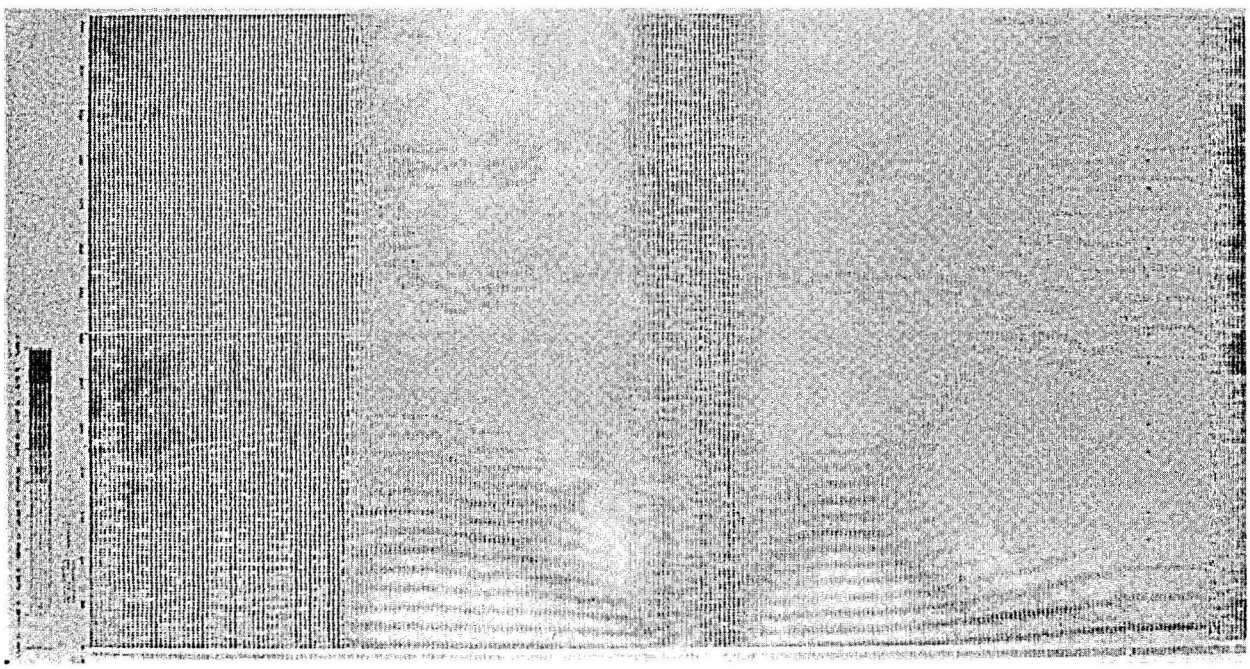


Abb.1b, 'HAWAII', 2000 bit/sek

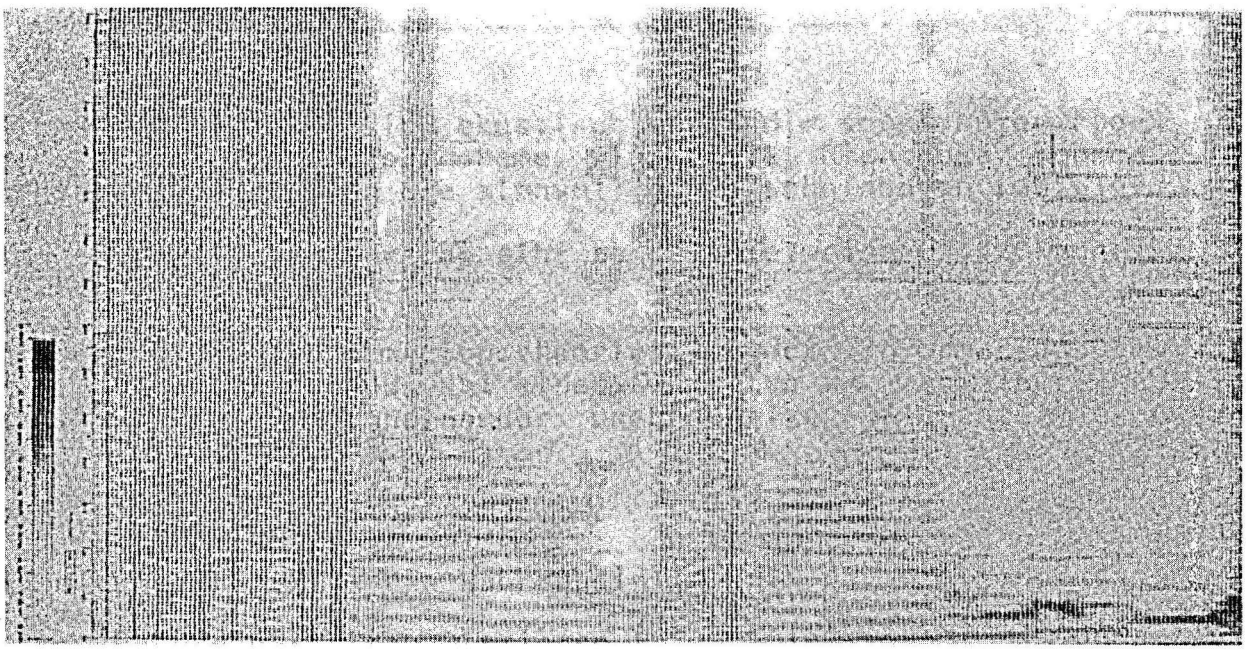


Abb.1c, 'HAWAII', 1000 bit/sek

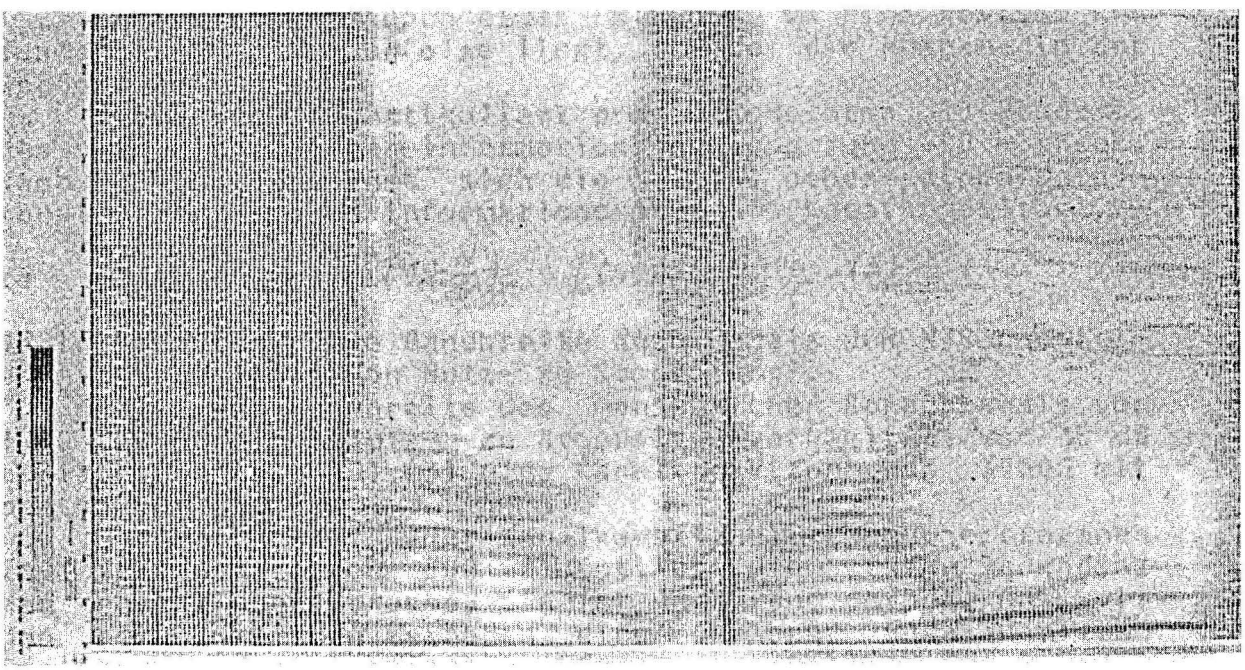


Abb.1d, 'HAWAII', 800 bit/sek

2. Theorie der Spracherzeugung und Sprachkompression

2.1. Informationsgehalt der Sprache

Die Sprache laesst sich akustisch durch die sogenannten Phoneme beschreiben. Die Phoneme, wie Vokale, Diphtonge, Konsonanten usw. stellen die sinnvoll unterscheidbaren Lautelemente der Sprache dar.
In der englischen Sprache gibt es beispielweise ca 42 Phoneme.

Nach der Informationstheorie ist der Informationsgehalt, der mit der Auswahl eines Elementes x_i aus einer Anzahl diskreter, voneinander unabhaengiger Elemente verknuepft ist:

$$I = -ld\{p(x_i)\} \quad [bit] \qquad (1)$$

Dabei stellt $p(x_i)$ die Wahrscheinlichkeit dar, mit der das Element x_i auftritt.
Eine Aussage ueber den mittleren Informationsgehalt der Phoneme stellt die Entropie dar:

$$H(X) = \sum_i p(x_i) \cdot ld\{p(x_i)\} \quad [bit] \qquad (2)$$

Sie betraegt fuer die englische Sprache 4.9 bit. Dieser Wert ist in Wirklichkeit noch etwas kleiner, da eine gewisse Redundanz in der Reihenfolge liegt, in der die Phoneme in der Sprache auftreten.

Ein Sprecher artikuliert pro Sekunde etwa 10 Phoneme. Das entspricht einer Informationsrate von rund 50 bit/sek. Nach Shannon berechnet sich die maximal ueber einen Datenkanal uebertragbare Informationsrate, die Kanalkapazitaet,

zu $$C = B \cdot ld\left(1+\frac{N}{S}\right) \quad [bit/sek] \qquad (3)$$

Darin bedeuten B die Bandbreite des Kanals und N/S das Leistungsverhaeltnis von Nutz- zu Stoersignal.

Bei einer Bandbreite des menschlichen Sprachkanals von 3000 Hz und einem Nutz- zu Stoersignalverhaeltnis von 30 dB ergibt sich nach Gl.(3) eine Kanalkapazitaet von 30000 bit pro sek.

Das Verhaeltnis der Kanalkapazitaet zur uebertragenen Informationsrate von 600 verdeutlicht die starke Redundanz der menschlichen Sprache. Sie ist der Grund dafuer, dass mit den herkoemmlichen Methoden eine nur sehr unoekonomische Speicherung der akustischen Information moeglich ist.

Die Redundanz der Sprache liegt erstens im physiologischen Aufbau des menschlichen Spracherzeugungssystems und in der besonderen Art begruendet, in der sich der Mensch geuebt

hat, seine Sprachorgane zu benutzen.
Zweitens liegt die Redundanz der Sprache in bestimmten Regeln, nach denen der Mensch Laute zu Worten und unter Hinzufuegung von Betonungen Worte zu sinnvollen Saetzen formt.

Der zweite Punkt kennzeichnet das Gebiet, das unter dem Schlagwort 'Speech Synthesis by Rule' bearbeitet wird.
Der Verfasser hat sich in dieser Dissertation mit der erstgenannten Ursache fuer die Sprachredundanz beschaeftigt und ein Spracherzeugungssystem entwickelt, das vom systemtheoretischen Standpunkt weitgehend die Eigenschaften des menschlichen Spracherzeugungssystems aufweist. Damit wird die Uebertragung der Information entbehrlich gemacht, die im physiologischen Aufbau und den typischen Bewegungen der Spracherzeugungsorgane des Menschen begruendet ist.

2.2 Aufbau und mathematische Beschreibung des menschlichen Spracherzeugungssystems

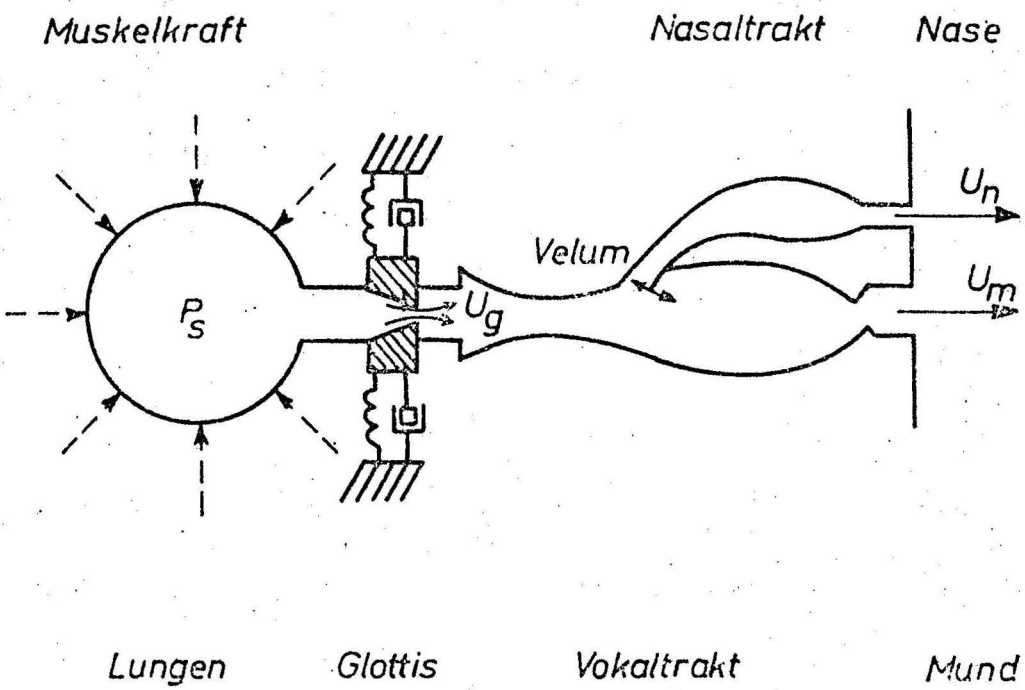


Abb.2, Schema des menschlichen Spracherzeugungssystems (/1/ S.24)

Die Abb.2 zeigt schematisch die wichtigsten Teile des menschlichen Spracherzeugungssystems.

Auf der linken Seite sind die Lungen als Luftreservoir dargestellt. In dem Luftreservoir wird durch die Muskelkraft des Brustkorbes der Druck P_S erzeugt. Der Druck verursacht einen Luftstrom, der die Stimmritze zu Schwingungen anregt.

Die Stimmritzen stellen einen mechanischen Resonator dar. Die Elemente dieses Resonators sind durch die Symbole fuer die Masse, die Federkraft und die Daempfung in der Abb.2 angedeutet.

Der zeitliche Verlauf der Volumendurchtrittsgeschwindigkeit an der Glottis u_G entspricht, bedingt durch die Schwingungen der Glottis, einer aequidistanten Pulsfolge.

Von der Stimmritze bis zur Mundoeffnung erstreckt sich der Vokaltrakt. Er hat bei einem erwachsenen Mann eine Laenge von ca 17 cm.
Der Querschnitt des Vokaltraktes ist i.a. nicht konstant. Er haengt ab von der jeweiligen Mundstellung, der Oeffnung der Lippen und der Lage der Zunge.

Der Vokaltrakt ist ein mechanischer Resonator, dessen Uebertragungseigenschaften durch Aenderung des jeweiligen Querschnittverlaufs und das Hinzuschalten des nasalen Trakts variiert werden koennen.

Ganz allgemein laesst sich eine Uebertragungsfunktion, die ja lt. Definition eine reelle Funktion ist, durch konjugiert komplexe Pol- und Nullstellenpaare bzw. einfach reelle

Pole und Nullstellen beschreiben. Unter Vernachlaessigung der einfach reellen Punkte und Nullstellen und unter der Bedingung

$$|H(s)|_{j\omega=0} = 1 \qquad (4)$$

ergibt sich die allgemeine Darstellung einer Uebertragungsfunktion nach Gl.(5):

$$H(s) = \prod_{i=1}^{n} \frac{s_i \cdot s_i^*}{(s-s_i)\cdot(s-s_i^*)} \cdot \prod_{j=1}^{m} \frac{(s-s_j)(s-s_j^*)}{s_j \cdot s_j^*} \qquad (5)$$

Ein konjugiert komplexes Polpaar der Form

$$H_p(s) = \frac{s_p \cdot s_p^*}{(s-s_p)\cdot(s-s_p^*)} \qquad (6)$$

wird Formant genannt.

Die Nullstellenpaare der Form

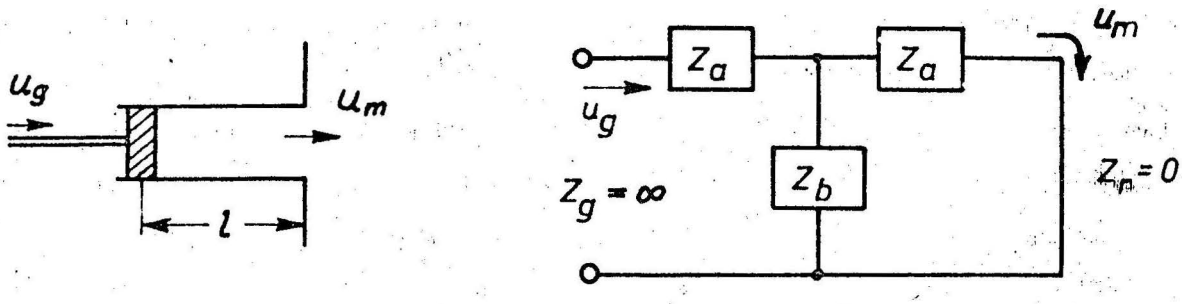
$$H_z(s) = \frac{(s-s_z)(s-s_z^*)}{s_z \cdot s_z^*} \qquad (7)$$

werden als Antiformanten bezeichnet.

Die Uebertragungsfunktion des Vokaltraktes laesst sich also allgemein durch Formanten und Antiformanten beschreiben.

Bei der Artikulation eines 'æ' verhaelt sich der Vokaltrakt naeherungsweise wie ein einseitig geschlossenes Rohr mit konstantem Querschnitt, das an seinem abgeschlossenen Ende, der Glottis, durch einen Generator 'konstanter Volumendurchtrittsgeschwindigkeit' angeregt wird.

Die akustische und die Netzwerkbeschreibung fuer diesen Fall sind in Abb.3 dargestellt (/2/ S.52).



$$Z_a = Z_0 \tanh \frac{\gamma \cdot l}{2}$$

$$Z_b = \frac{Z_0}{\sinh \gamma l} \qquad (8)$$

$$\gamma = \alpha + j\beta$$

Abb.3 Vereinfachtes Modell des Vokaltraktes

u_g stellt die Volumendurchtrittsgeschwindigkeit an der Glot-

tis und u_m die an der Mundoeffnung dar.
Der Innenwiderstand Z_g des Generators, der den Innenwiderstand der Glottis darstellt, ist unendlich gross und der Abstrahlungswiderstand Z_r an der Mundoeffnung vernachlaessigbar klein.
Dann ergibt sich als Uebertragungsfunktion des Vokaltraktes aus Gl.(8):

$$\frac{u_m}{u_g} = \frac{Z_b}{Z_a + Z_b} = \frac{1}{\cosh(\gamma l)} \qquad (9)$$

Diese Uebertragungsfunktion weist Pole fuer $\cosh(\gamma l) = 0$ auf.

$$\cosh(\gamma l) = 0 \rightsquigarrow \gamma l = \pm j(2n-1)\frac{\pi}{2} \quad n = 1,2,\ldots \qquad (10)$$

Aus $\gamma = \alpha + j\beta$ und $\beta = \frac{\omega}{c}$ und der Voraussetzung einer kleinen Daempfung ergeben sich die Lagen der Resonanzfrequenzen naeherungsweise zu :

$$s_n \approx -\alpha c \pm j\,\frac{(2n-1)\pi c}{2l} \qquad n = 1,2,\ldots \qquad (11)$$

Da die Resonanzfrequenzen in der komplexen Ebene konjugiert komplexe Werte darstellen, kann man die Uebertragungsfunktion des Vokaltraktes unter der Voraussetzung nach Gl.(4) auch schreiben:

$$H(s) = \frac{u_m(s)}{u_g(s)} = \prod_{n=1}^{\infty} \frac{s_n \cdot s_n^*}{(s-s_n)\cdot(s-s_n^*)} \qquad (12)$$

Es zeigt sich, dass in diesem Falle keine Antiformanten auftreten.
Die Resonanzfrequenzen berechnen sich aus Gl.(11) zu

$$\omega = \pm(2n-1)\frac{\pi c}{2l} \rightsquigarrow f = \pm\frac{(2n-1)\cdot c}{4l} \qquad (13)$$

Setzt man, wie oben erwaehnt wurde, l=17 cm und c=340 m/s, ergeben sich als Resonanzfrequenzen die Werte:

500 Hz, 1500 Hz, 2500 Hz, 3500 Hz, 4500 Hz, ...

Die Uebertragungsfunktion des Vokaltraktes wird durch die Mundbewegungen veraendert. Dadurch weichen die Werte der Formantfrequenzen von den oben berechneten nach beiden Seiten mehr oder weniger ab.
Es ergeben sich dabei fuer die ersten drei Formanten ungefaehr die folgenden Frequenzbereiche:

f1: 200 - 900 Hz
f2: 550 - 2700 Hz
f3: 1100 - 2950 Hz

Der Vokaltrakt stellt nur einen Teil des mechanischen Filters des Spracherzeugungssystems dar. Parallel dazu liegt der nasale Trakt, der durch das Velum zu- oder abgeschaltet werden kann.

Beim Vorhandensein nasaler Laute wird der Schall, wie aus Abb.2 ersichtlich ist, zusaetzlich zur Schnelle u_m an der Mundoeffnung durch die Schnelle u_n an der Nasenoeffnung

abgestrahlt.
Der Strahlungswiderstand an der Mundoeffnung wurde in Abb.3, um eine einfache Abschaetzung der Eigenfrequenzen des Vokaltraktes durchfuehren zu koennen, vernachlaessigt.

Bei Wellenlaengen, die gross gegenueber den Abmessungen des Vokaltraktes sind, kann man davon ausgehen, dass die Schnelle an Mund- und Nasenoeffnung inphase ist. Geht man so weit, dass man die Schallabstrahlung durch eine schwingende Kugel darstellt, so ergibt sich als normalisierte Impedanz (/2/ S.33)

$$\underline{z} = \frac{jka}{1+jka} \qquad k = \frac{\omega}{c} \qquad (14)$$

a stellt dabei den Kugelradius dar.

Dieser Wert kann als erste Naeherung fuer den Strahlungswiderstand des menschlichen Spracherzeugungssystems angenommen werden.
Gl.(14) laesst sich in der komplexen Ebene durch einen Pol auf dem negativen Teil der reellen Achse und eine Nullstelle im Ursprung darstellen.

2.3 Lauterzeugung

Die Laute, die mit dem menschlichen Spracherzeugungssystem artikuliert werden koennen, lassen sich aufgrund ihrer unterschiedlichen Entstehungsweise in Kategorien einteilen. An dieser Stelle sollen die 4 Lautkategorien Vokale, Nasale, Stimmlose Laute und Plosive besprochen werden, die nach Meinung des Verfassers in der deutschen und englischen Sprache am wichtigsten sind.

Die Sprache laesst sich noch feiner durch weitere Aufteilung der Lautkategorien unterteilen. Hierzu siehe B. GOLD und C.M. RADER /4/ S.132.

Vokale

Beispiele fuer Vokale sind : a, ai, æ, ɛ, ʌ, i, ɔ, ou

Der gleichfoermige Luftstrom, der von den Lungen herruehrt, wird durch die Stimmritzen in eine quasiperiodische Pulsfolge verwandelt. Der zeitliche Verlauf der Schnelle an der Glottis verlaeuft, wie Abb.4 zeigt, dreieckfoermig. Das Spektrum der Pulsfolge faellt daher mit ca 12 dB/Okt ab.

Dieses Spektrum, das noch bis zu sehr hohen Frequenzen betraechtliche Amplitudenwerte aufweist, wird durch das nachfolgende mechanische Filter, dem Vokaltrakt, gefiltert.

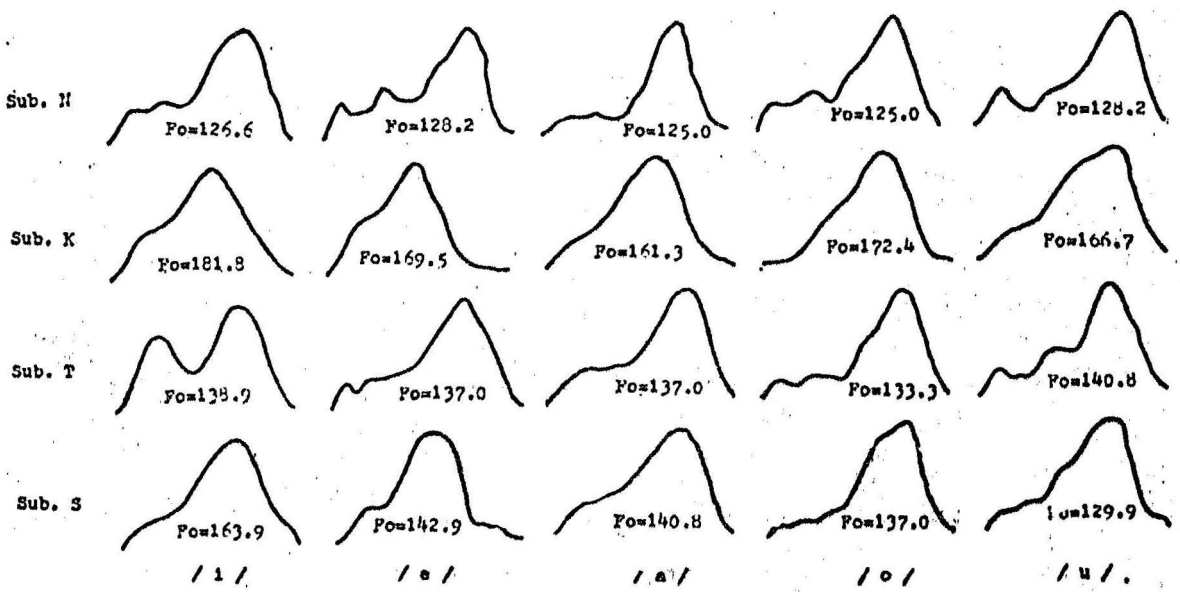


Abb.4, Pulsverlaeufe an der Glottis /3/

Da der Mund bei der Erzeugung von Vokalen mehr oder weniger geoeffnet ist und der nasale Trakt durch das Velum abgetrennt ist, sind in erster Naeherung die Voraussetzungen erfuellt, die zur vereinfachten Darstellung des Vokaltraktes nach Abb.3 fuehrten. Die Uebertragungsfunktion kann daher

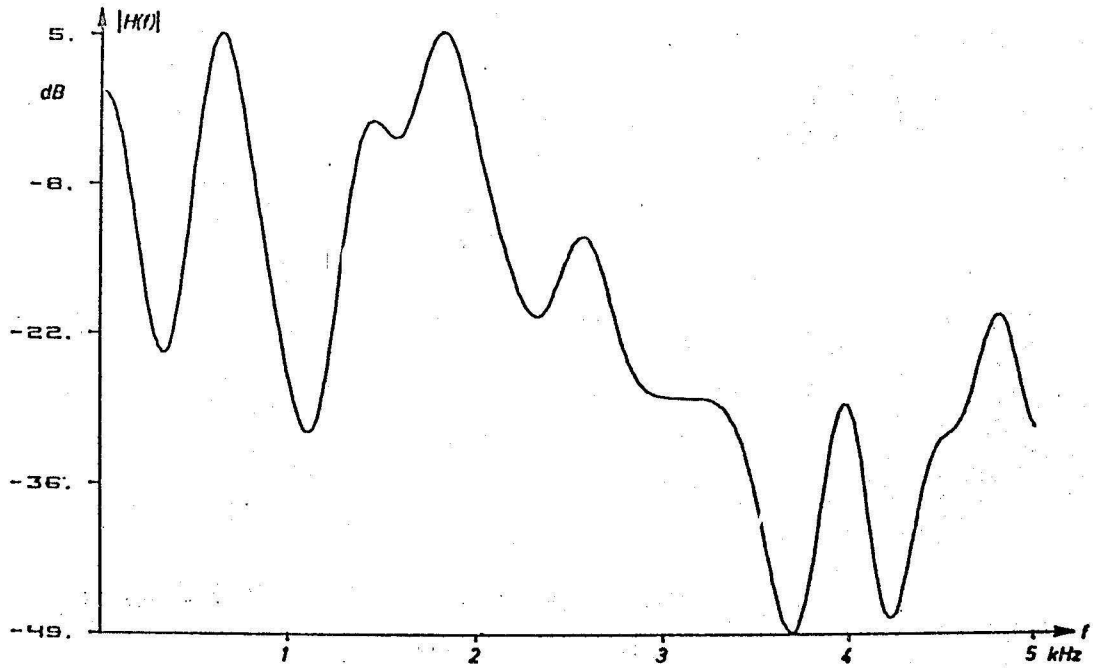


Abb.5, Gefiltertes Kurzzeitspektrum eines Vokals

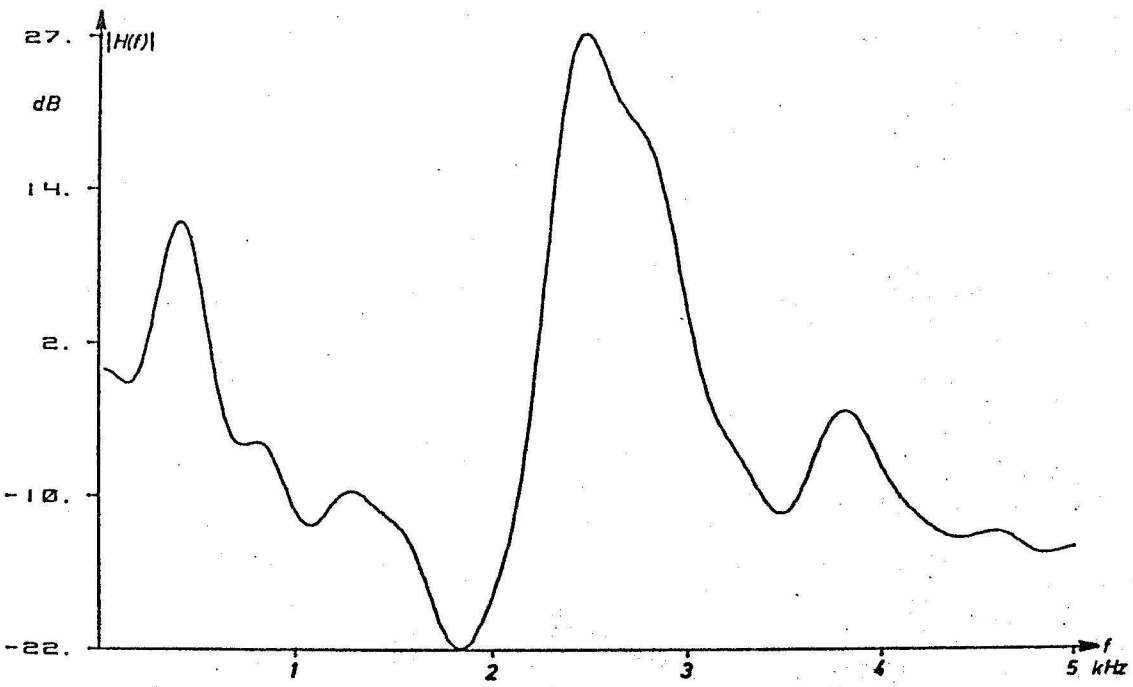


Abb.6, Gefiltertes Kurzzeitspektrum eines Nasals

durch die Gl.(12) beschrieben werden. Das bedeutet, dass bei der Erzeugung von Vokalen der Vokaltrakt alleine durch For-

manten charakterisiert werden kann.

Messungen haben gezeigt, dass drei bis sechs Formanten im Sprachspektrum von Vokalen gefunden werden koennen. Die Abb.5 zeigt das gefilterte Kurzzeitspektrum eines Vokals.

Nasale

Beispiele fuer Nasale sind: n, m, η

Bei der Erzeugung nasaler Laute wird der Trakt am Velum ebenfalls von der quasiperiodischen Pulsfolge der Glottis gespeist. Die Voraussetzungen, die unter 2.2 zur Berechnung der Uebertragungsfunktion aus einem vereinfachten Modell des mechanischen Sprachfilters zur Gl.(12) fuehrten, sind hier nicht mehr erfuellt.

Durch das Hinzuschalten des nasalen Traktes zum Vokaltrakt tritt eine Kopplung der beiden Trakte auf, die sich im Kurzzeitspektrum eines Nasals durch das Auftreten eines Antiformanten bemerkbar macht, wie aus Abb.6 ersichtlich ist.

Stimmlose Laute

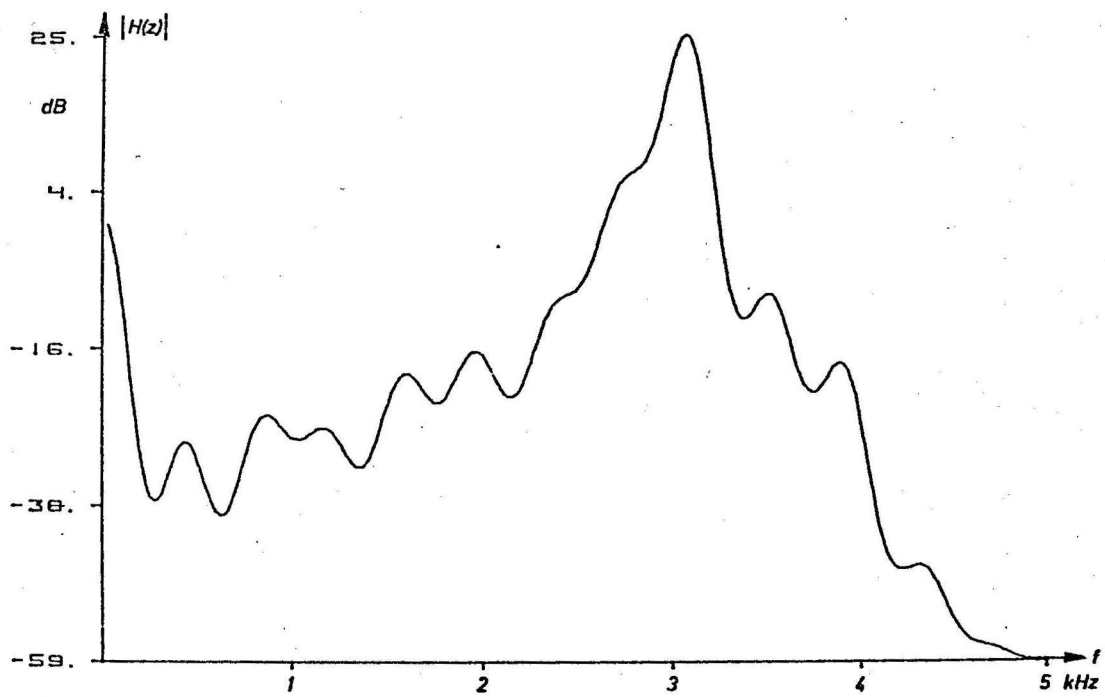


Abb.7, Gefiltertes Kurzzeitspektrum eines stimmlosen Lautes

Beispiele fuer stimmlose Laute sind: $s, z, \int, t\int, d\int, \theta, \beta, f$

Die stimmlosen Laute werden dadurch erzeugt, dass der Luftstrom aus den Lungen durch eine Konstriktion im Vokaltrakt gepresst wird. Die Turbulenz, die dabei an der Konstriktion entsteht, hoert sich akustisch wie Rauschen an und hat auch das typische breitbandige Rauschspektrum.

Der Schwerpunkt der Energie im Spektrum liegt meist im Bereich ueber 3000 Hz.

Die Abb.7 zeigt das gefilterte Kurzzeitspektrum eines stimmlosen Lautes. Wie die Abbildung zeigt, ist keine vergleichbare Formantstruktur, wie bei Vokalen erkennbar.

Das Verschwinden der Formantstruktur laesst sich dadurch erklaeren, dass der Vokaltrakt nicht, wie unter 2.2 angenommen wurde, von der Glottis bis zu den Lippen geoeffnet ist, sondern einen Verengung aufweist. Die Turbulenz an dieser Verengung, die die akustische Quelle zur Anregung des Vokaltraktes ist, befindet sich auch nicht, wie unter 2.2 angenommen wurde, am Ende des Traktes. Bei der Erzeugung des Lautes Θ liegt die Konstriktion beispielsweise unmittelbar hinter den Zaehnen und beim f an den Lippen.

Der Vokaltrakt laesst sich bei stimmlosen Lauten ausreichend durch zwei Formanten und einen Antiformanten beschreiben.

Plosive

Beispiele fuer Plosive sind: p, t, k, b, d, g

Bei der Erzeugung von Plosiven wird im Vokaltrakt kurzzeitig ein Verschluss erzeugt. Hinter der Verschlussstelle bildet sich ein hoher Druck aus, der durch eine abrupte Bewegung der Artikulationsorgane ploetzlich entweichen kann.

Die akustische Quelle zur Lauterzeugung der Plosive laesst sich am besten durch einen Generator beschreiben, der einen einzelnen Puls abgibt. Der Ort des Verschlusses und damit der akustischen Quelle liegt nicht, wie bei den Vokalen, an der Stimmritze, sondern beispielsweise bei der Erzeugung des p an den Lippen, des t am Gaumen und des k im Rachenraum.

Der Vokaltrakt laesst sich bei der Erzeugung von Plosiven in ausreichendem Masse durch zwei Formanten und einen Antiformanten beschreiben.

Charakteristisch fuer einen Plosiv ist ein auffaelliges Maximum im Amplitudenverlauf, wie aus Abb.8 ersichtlich wird.

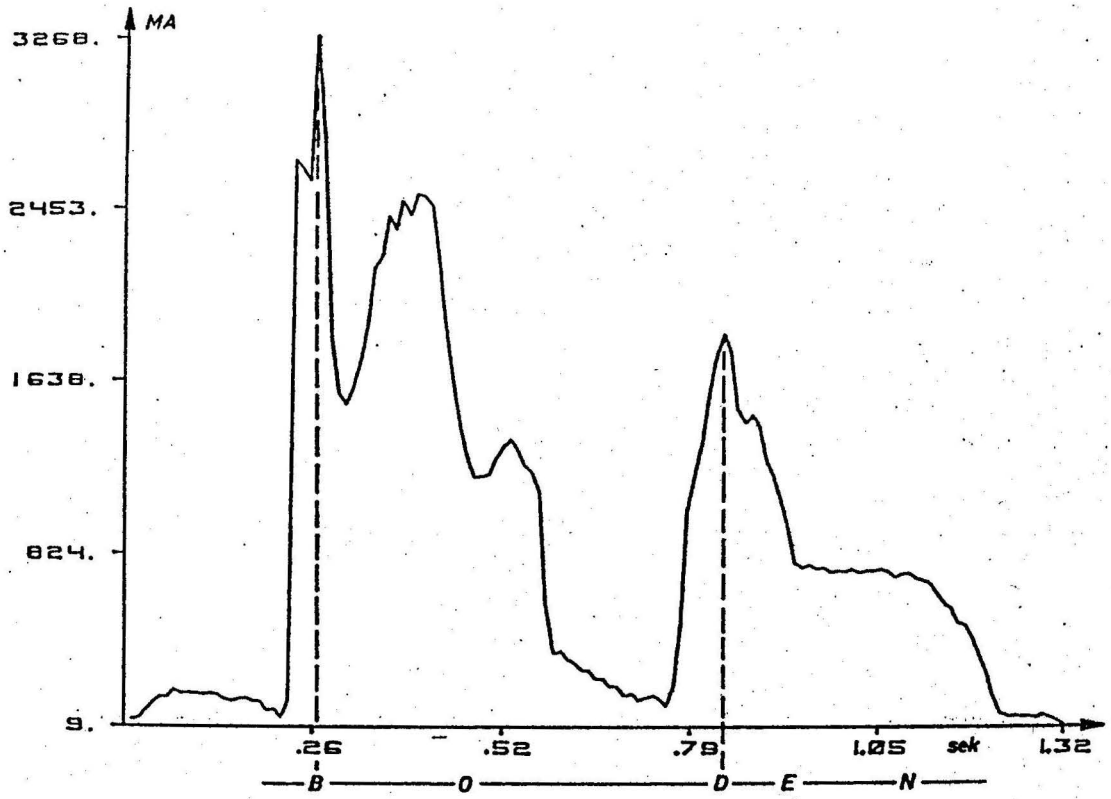


Abb.8, Amplitudenverlauf von BODEN

2.4 Kuenstliche Spracherzeugung

Es gibt die verschiedensten Methoden Sprache synthetisch zu erzeugen. Viele Verfahren sind dadurch entstanden, dass man lediglich versucht hat, die Bandbreite eines Uebertragungskanals zu verringern, beispielsweise, um mehrere Telephongespraeche ueber eine Leitung schicken zu koennen.

Eines dieser Verfahren ist bekannt unter dem Namen'Predictive Coding'. Durch einen Prediktor wird der Verlauf der Sprachzeitfunktion vorhergesagt und durch einen Korrektor die Abweichung von der Voraussage uebertragen. Wenn der Prediktor ideal ist, werden vom Korrektor nur noch Werte uebertragen, die nicht mehr miteinander korreliert sind und damit ist die Nachrichtenuebertragung von jeglicher Redundanz der Sprache befreit.

Eine weitere Moeglichkeit der Sprachbandkompression besteht darin, dass man mit einer Filterbank das Sprachsignal zunaechst in eine grosse Anzahl benachbarter Frequenzbaender aufteilt. Die fast sinusfoermig verlaufenden Ausgangssignale der Bandpaesse werden in ihrer Frequenz halbiert und wieder zu einem Gesamtsignal zusammengefasst, das jetzt nur noch die halbe Bandbreite zur Uebertragung braucht.

Auf der Empfaengerseite muss das Signal in derselben Weise wieder aufgespalten und durch Frequenzmultiplikatoren auf das Originalsignal zurueckgefuehrt werden.

Eine andere Moeglichkeit der Sprachsynthese besteht darin, dass man zunaechst versucht, den Verlauf der Sprachzeitfunktion in geeignete Segmente zu zerlegen. Anschliessend wird die Zeitfunktion innerhalb eines Segmentes durch andere Funktionen, z.B. orthogonale Funktionen approximiert.

Der Verfasser hat in dem Zusammenhang versucht, die Sprachsegmente aus Gaussfunktionen zusammenzusetzen.

Die weitaus groesste Bedeutung in der Literatur hat die Methode der Sprachsynthese, bei der das menschliche Spracherzeugungssystem auf systemtheoretischer Basis simuliert wird.

Wie aus 2.2 ersichtlich ist, besteht das Spracherzeugungssystem des Menschen aus drei Grundelementen:

1. Die akustischen Quelle

Sie hat den Zeitverlauf q(t) und das Spektrum:

$$Q(s) = \mathcal{L}[q(t)] \qquad (15)$$

Die Quelle gibt bei der Erzeugung von Vokalen und Nasalen eine quasiperiodische Pulsfolge ab. Die Frequenz liegt bei einem erwachsenen, maennlichen Sprecher im Bereich von 80 - 200 Hz. Das Spektrum der Pulsfolge weist einen Abfall von 12 dB/Okt auf.
Bei der Erzeugung stimmloser Laute wird die Quelle

durch einen Rauschgenerator dargestellt und bei der Erzeugung von Plosiven gibt sie einen einzelnen Puls ab.

2. Der Vokaltrakt

Der Vokaltrakt weist die Impulsantwort h(t) und die folgende Uebertragungsfunktion auf:

$$H(s) = \mathcal{L}[h(t)] \qquad (16)$$

Das Ohr ist relativ unempfindlich gegenueber Phasenlagen, und daher ist nur der Betrag der Uebertragungsfunktion von Bedeutung.
Der Vokaltrakt hat bei der Erzeugung von Vokalen eine reine Formantstruktur und wird bei stimmlosen Lauten und bei Plosiven ausreichend durch zwei Formanten und einen Antiformanten beschrieben.

3. Die Abstrahlung

Die Abstrahlung wird durch den Abstrahlungswiderstand nach Gl.(14) beruecksichtigt. Er entspricht einem Filter mit der Impulsantwort r(t) und der Uebertragungsfunktion:

$$R(s) = \mathcal{L}[r(t)] \qquad (17)$$

Der Druckverlauf im Schallfeld eines Sprechers berechnet sich zu:

$$p(t)=q(t)*h(t)*r(t) \qquad (18)$$

Die Sterne deuten dabei die Faltung an.
Aus Gl.(15), Gl.(16) und Gl.(17) folgt mit Gl.(18):

$$P(s)=Q(s)\cdot H(s)\cdot R(s) \qquad (19)$$

In Kapitel 3 werden u.a. mehrere Sprachkompressionssysteme betrachtet, die in ihrer Funktionsweise auf Gl.(18) und Gl.(19) aufbauen.

Die Sprachkompressionssysteme, auch Vocoder genannt, bestehen alle aus einem Analyse- und einem Syntheseteil. Im Analyseteil werden die Parameter aus der Sprache extrahiert, die mit erheblich geringerer Bandbreite gegenueber der Originalsprache uebertragen oder mit wesentlich geringerem Speicherplatzbedarf abgespeichert werden koennen.

Aus diesen Parametern wird im Syntheseteil, natuerlich unter einer unvermeidlichen Qualitaetseinbusse, die Originalsprache wieder rekonstruiert.

Beispiele fuer Vocoderarten, die auf Gl.(18) und Gl.(19) basieren sind der Autokorrelationsvocoder, der Kanalvocoder und der Formantvocoder.

Eine weitere Moeglichkeit der kuenstlichen Spracherzeugung liegt in den sog. artikulatorischen Vocodern.

Die Parameter, die eine derartige Sprachsynthese steu-

ern, beschreiben den Querschnittverlauf des Vokaltraktes. Beispielweise werden nach FLANAGAN (/1/ S.35) die folgenden sieben artikulatorischen Parameter verwendet:

2 Koordinaten zur Beschreibung der Lippenkonfiguration
2 Koordinaten geben die Lage der Zungenspitze an
2 Koordinaten lokalisieren die Lage des Zungenkoerpers
1 Koordinate beschreibt die Lage des Velums.

Die Wellenausbreitung im Vokaltrakt wird naeherungsweise durch die WEBSTER'sche Differentialgleichung beschrieben:

$$\frac{1}{A(x)} \frac{\partial}{\partial x} \left[A(x) \frac{\partial p}{\partial x} \right] = \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} \qquad (20)$$

Dabei bedeuten:
p(x,t) = Schalldruck als Funktion der Laengskoordinate x
A(x) = Querschnittsverlauf des Vokaltraktes
c = Schallgeschwindigkeit.

Mit Hilfe der Randbedingungen, die durch die artikulatorischen Parameter gegeben sind, werden die ersten drei Eigenfrequenzen des Vokaltraktes berechnet. Die ersten drei Eigenfrequenzen entsprechen den ersten drei Formanten. Die eigentliche Synthese erfolgt dann mit einem Formantvocoder.

Dieser Vocodertyp, der gegenueber dem Formantvocoder vor allem eine hoehere Stufe der Formantanalyse darstellt, wird vom Verfasser in dieser Arbeit nicht weiter behandelt.

3. Verschiedene Methoden der synthetischen Spracherzeugung

In 2.4 wurde auf die verschiedenen Moeglichkeiten einer kuenstlichen Spracherzeugung bereits hingewiesen. Es gibt in der Literatur eine Vielzahl von Vocoderarten, die fuer die verschiedensten Verwendungszwecke entwickelt worden sind. Aus dieser Vielzahl soll eine Auswahl derjenigen Vocoderprinzipien getroffen werden, die fuer die vorliegende Aufgabenstellung am geeignetsten erscheinen. Die Anforderungen, die das gesuchte Vocodersystem erfuellen muss, sind:

1. Das Vocodersystem soll einen hohen Kompressionsfaktor bei guter Verstaendlichkeit der synthetischen Sprache aufweisen.
2. Es soll lediglich ein Standardtext im Digitalrechner abgespeichert werden, d.h. ein endlicher Wortvorrat wird durch eine einmalige Analyse geschaffen.
3. Der Aufbau des Syntheseteils muss so beschaffen sein, dass eine hardwaremaessige Vorrichtung aufgebaut werden kann, die eine Spracherzeugung in Realzeit ermoeglicht.

Der Sprachkompressionsfaktor berechnet sich aus einem Bezugswert, dividiert durch die Bitrate, die zur Codierung der Ausgangsparameter des Analyseteils benoetigt wird. Der Bezugswert ergibt sich aus der Annahme, dass die Originalsprache mit 10 kHz abgetastet und mit 8 bit linear quantisiert wird zu 80000 bit/sek.

Wenn man Sprachkompressionsfaktoren groesser als 20 fordert, so scheiden fuer die hier durchgefuehrte Auswahl der Voice-Excited-Vocoder (/5/,/8/S.643) und der Phasenvocoder /6/ aus. Auch Vocoder, die auf dem Prinzip der Frequenzdivision (/7/ S.730) arbeiten und viele andere, die beispielsweise entwickelt worden waren, um Sprache von 10 kHz Bandbreite ueber einen Telephonkanal von 3kHz Bandbreite zu uebertragen, koennen hier dann nicht beruecksichtigt werden.

Ueber die Rechenzeit, die fuer die Analyse notwendig ist, wurde keine Aussage gemacht. Sie spielt primaer keine Rolle. Im Interesse der Wirtschaftlichkeit des Analyseverfahrens muss jedoch darauf geachtet werden, ein optimales Verhaeltnis der erzielten Sprachqualitaet zur aufgewendeten Rechenzeit zu erzielen.

3.1 Spracherzeugung aus Gaussfunktionen (/13/,/14/,/15/)

Im folgenden wird eine Moeglichkeit der Sprachanalyse- synthese beschrieben, die sich aus der Betrachtung des Zeitverlaufs einer Sprachschwingung ergibt.

Prinzip

Der Zeitverlauf der Sprache moege in Segmente endlicher Laenge zerlegt werden. Man kann dann die eindeutige Zuordnung der Spannungswerte am Ausgang eines Mikrophons zu den Zeitwerten mathematisch als Funktion f(t) auffassen und versuchen, diese durch einen Satz anderer Funktionen g_i anzunaehern, so dass sich

$$f(t) = \sum_{i=1}^{n} a_i g_i (t-t_i) \qquad (21)$$

ergibt. Die Guete der Approximation wird beeinflusst durch guenstige Segmentierung und die Wahl geeigneter Funktionen g_i

Zunaechst ein Wort zu 'Wahl geeigneter Funktionen'. Eine besonders rasche Approximation ist dann zu erwarten, wenn die Funktionen g_i bereits von sich aus einen aehnlichen Verlauf wie eine Sprachschwingung haben. Sie muessen dabei in dem betrachteten Bereich dem Betrag nach endlich bleiben, mehrere Nullstellen aufweisen und fuer grosse Argumente gegen Null konvergieren. Schliesslich, wenn man an eine hardwaremaessige Synthese in Analogtechnik denkt, soll sich ihr Zeitverlauf moeglichst einfach durch eine Rechenschaltung auf dem Analogrechner darstellen lassen.
Diesen Anforderungen genuegen beispielsweise die Gaussschen Funktionen.

Die Anregung zur Wahl der Gaussfunktionen geht auf eine Veroeffentlichung von J.A. HOWARD und R.C. WOOD /11/ zurueck. In der vorliegenden Arbeit wurde ein von HOWARD und WOOD unabhaengiger Weg fuer die Darstellung stimmhafter Laute durch Gausssche Funktionen beschritten /13/, der im folgenden beschrieben werden soll:

Die Gaussschen Funktionen n-ter Ordnung sind die Loesungen der Differentialgleichung:

$$\frac{d^2 G_n}{dt^2} + t\,\frac{dG_n}{dt} + (n+1)\cdot G_n = 0 \qquad (22)$$

Fuer das Analyseprogramm wurden Gaussfunktionen bis zur 10. Ordnung vorgesehen. Die Praxis zeigt aber, dass nur Gaussfunktionen bis zur 5. Ordnung benoetigt werden.

Um eine Vorstellung vom Verlauf des Graphen der Funktionen zu geben, sind in Abb.9, Abb.10 und Abb.11 Gaussfunktionen von nullter bis achter Ordnung graphisch dargestellt worden. Dabei ist zu beachten, dass die Gaussfunktionen gerader Ordnung sich als gerade Funktionen und die ungerader Ordnung sich als ungerade Funktionen in den Bereich negati-

ver Argumente. fortsetzen.

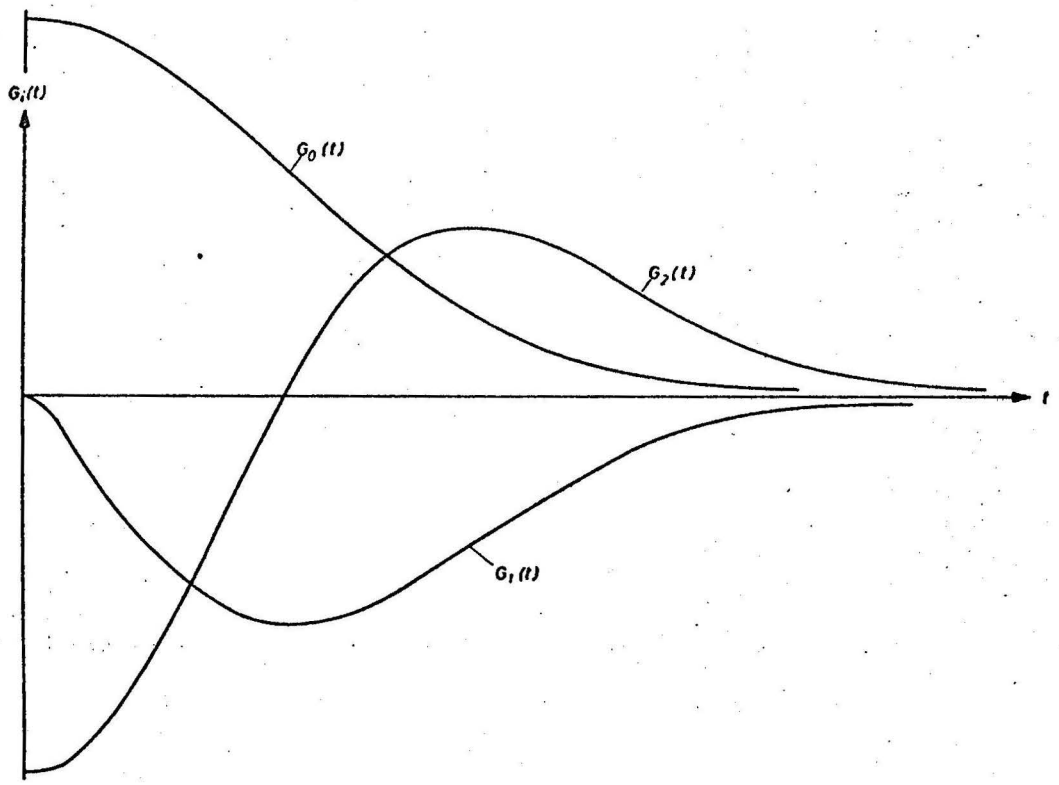


Abb.9, Gaussfuntionen 0-ter bis 2-ter Ordnung

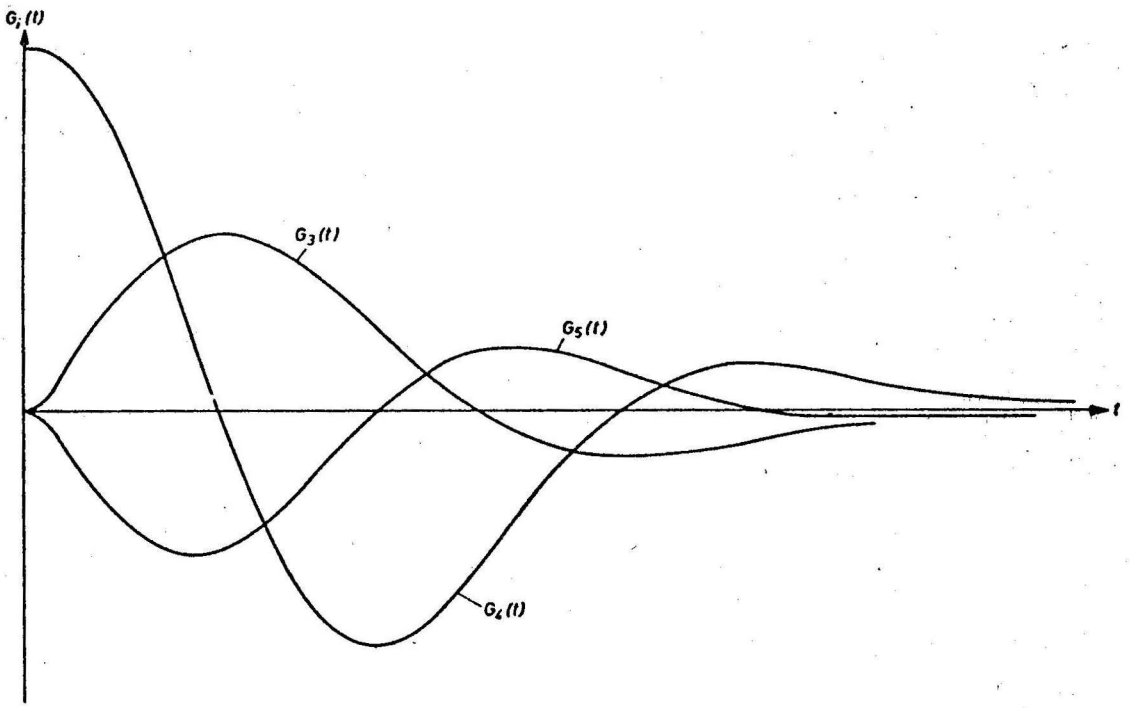


Abb.10, Gaussfunktionen 3-ter bis 5-ter Ordnung

ver Argumente. fortsetzen.

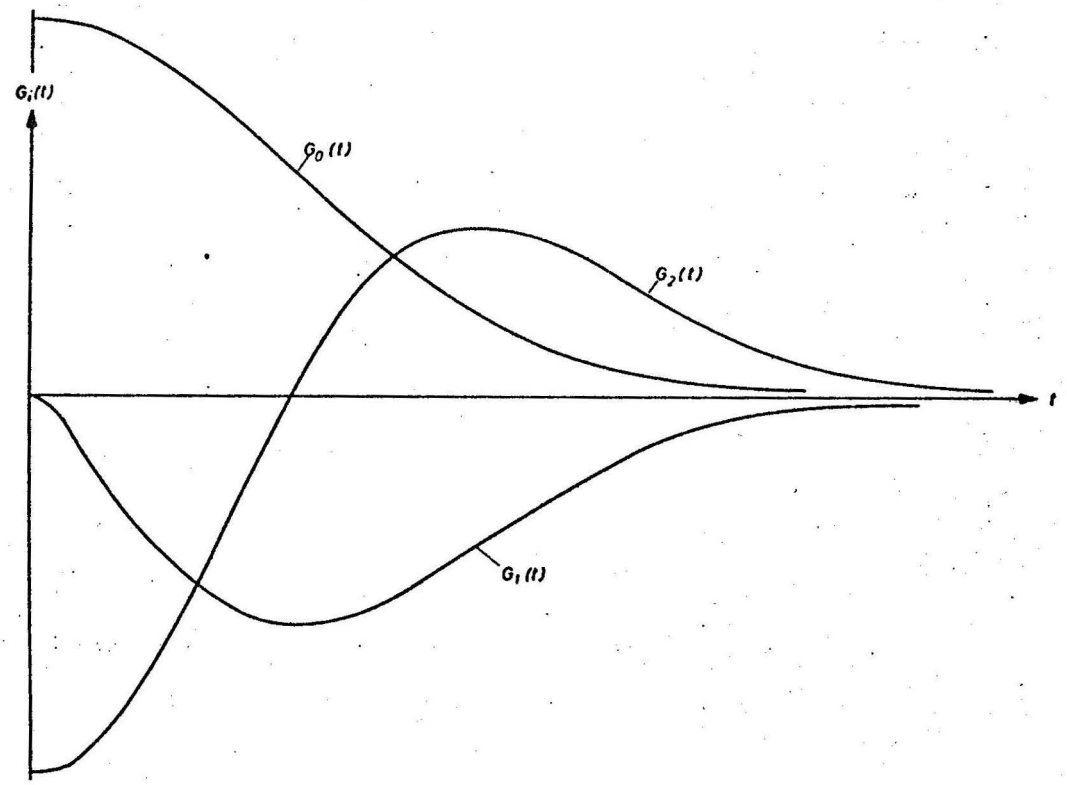


Abb.9, Gaussfuntionen 0-ter bis 2-ter Ordnung

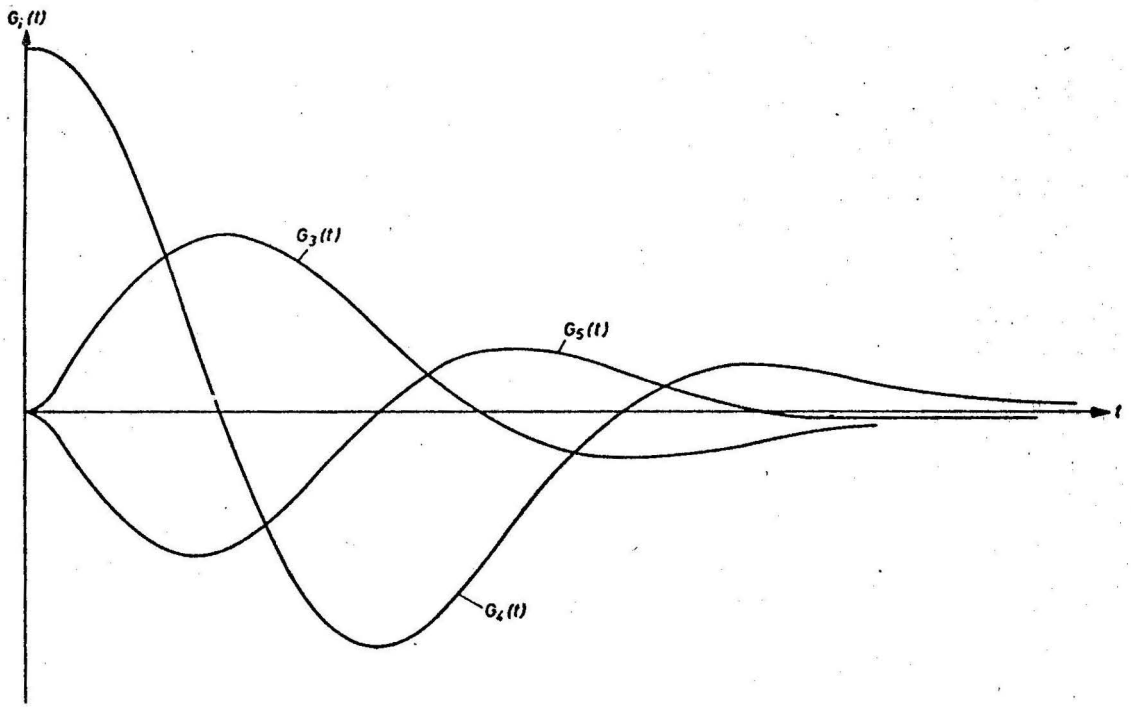


Abb.10, Gaussfunktionen 3-ter bis 5-ter Ordnung

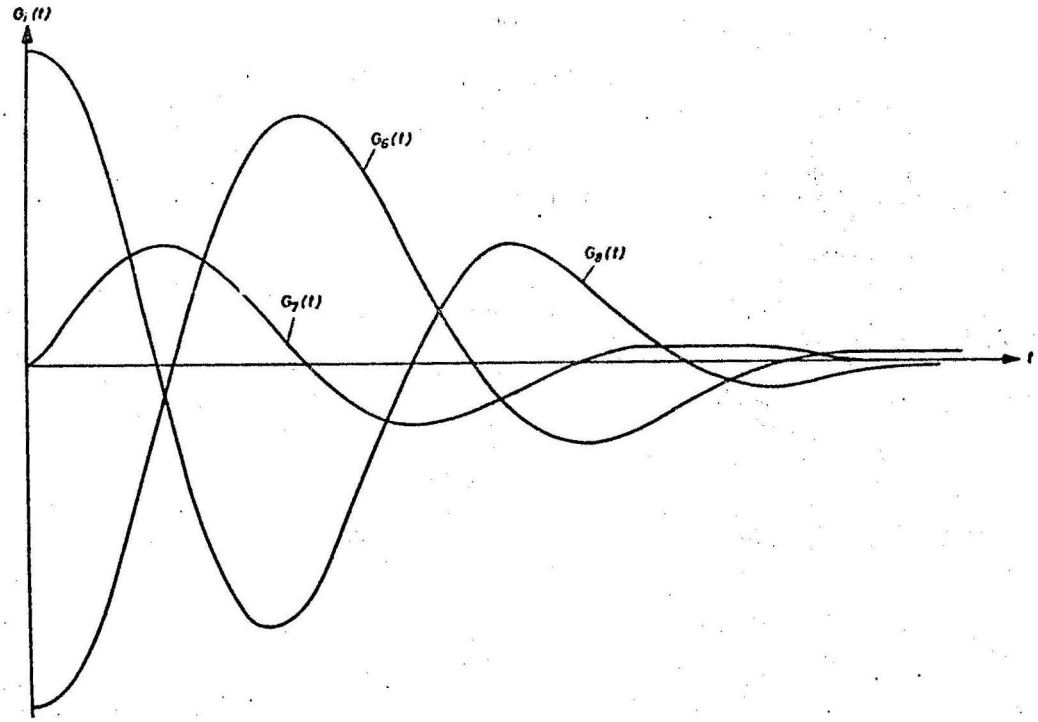


Abb.11, Gaussfunktionen 6-ter bis 8-ter Ordnung

Segmentierung

Der erste Schritt in der Analyse ist die Aufspaltung der Zeitfunktion in die Bereiche, in denen die eigentliche Analyse stattfinden soll. Die Peak-Struktur stimmhafter Laute kommt der Segmentierung sehr entgegen, wenn man als Laenge eines Segments gerade eine Pitchperiode der Sprache nimmt. Es wurde daher zunaechst eine Markierung der Sprache nach D.R. REDDY /12/ vorgenommen.

Der Abstand zweier aufeinanderfolgender 'Significant Maximum Peaks', der im folgenden mit SMP abgekuerzt werden soll, entspricht gerade einer Pitchperiode. Der Bereich von einem SMP bis zum naechsten eignet sich jedoch nicht zur Entwicklung nach Gaussfunktionen, da die Gaussfunktionen gerade und ungerade Funktionen sind, die zu positiven und negativen Argumenten abfallen. Man muss deshalb auch den Analysebereich so waehlen, dass das absolute Maximum, das hier durch den SMP gekennzeichnet ist, ebenfalls mehr in die Mitte des Analysebereiches verlegt wird.

Die genaue Bestimmung des Analyseintervalls geschieht folgendermassen: Es wird vom SMP zu positiven (bzw. negativen) Argumenten das Maximum M gesucht, das kleiner als das folgende Maximum M^+ ist. Die Nullstelle vor dem Maximum M wurde als Grenze des Analysebereiches gewaehlt.

Abb.12 verdeutlicht die Segmentierung.

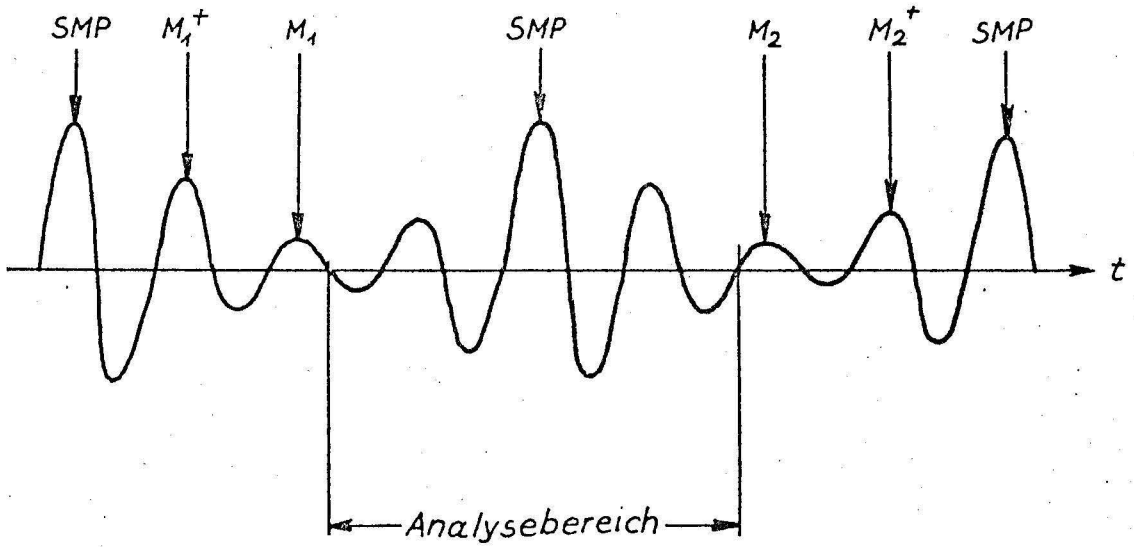


Abb.12, Segmentierung

Nullpunktsbestimmung und Entwicklung nach Gaussfunktionen

Fuer die Entwicklung nach Gaussfunktionen muss in dem angegebenen Analysebereich der Nullpunkt der Gaussfunktionen geeignet definiert werden. Die genaue Wahl des Nullpunktes haengt von der Umgebung des SMP ab.

In Abb.13 gilt folgende Bedingung fuer die dem SMP benachbarten Minima:

$$0.8 \leq \frac{a_1}{a_2} \leq 1.2$$

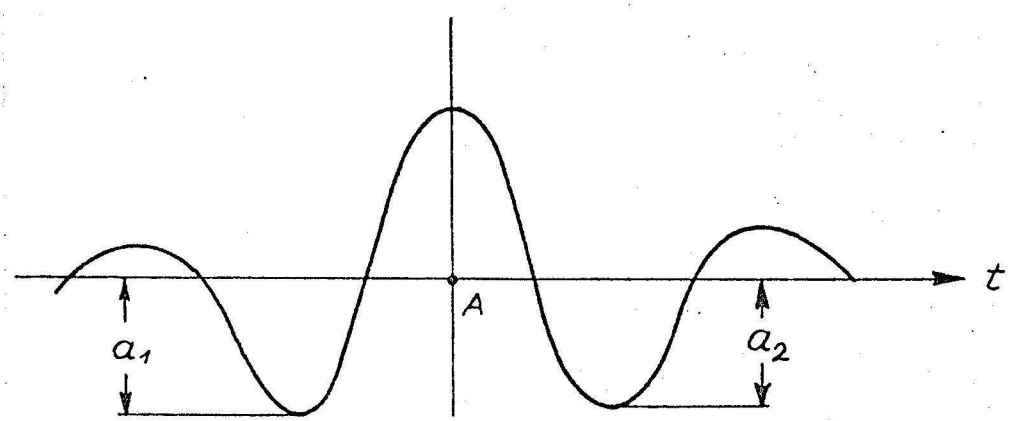


Abb.13, Nullpunktslage fuer die Entwicklung nach geraden Gaussfunktionen

In dem Fall wird der Nullpunkt zweckmaessigerweise nach A gelegt, d.h. an die Stelle des SMP, und die Entwicklung nach geraden Gaussfunktionen durchgefuehrt.

Abb.14 zeigt den Fall, in dem die in Abb.13 erwaehnte Bedingung nicht erfuellt ist. Dann wird der Nullpunkt fuer

die Gaussfunktion nach B verlegt und die Entwicklung nach ungeraden Funktionen durchgefuehrt.

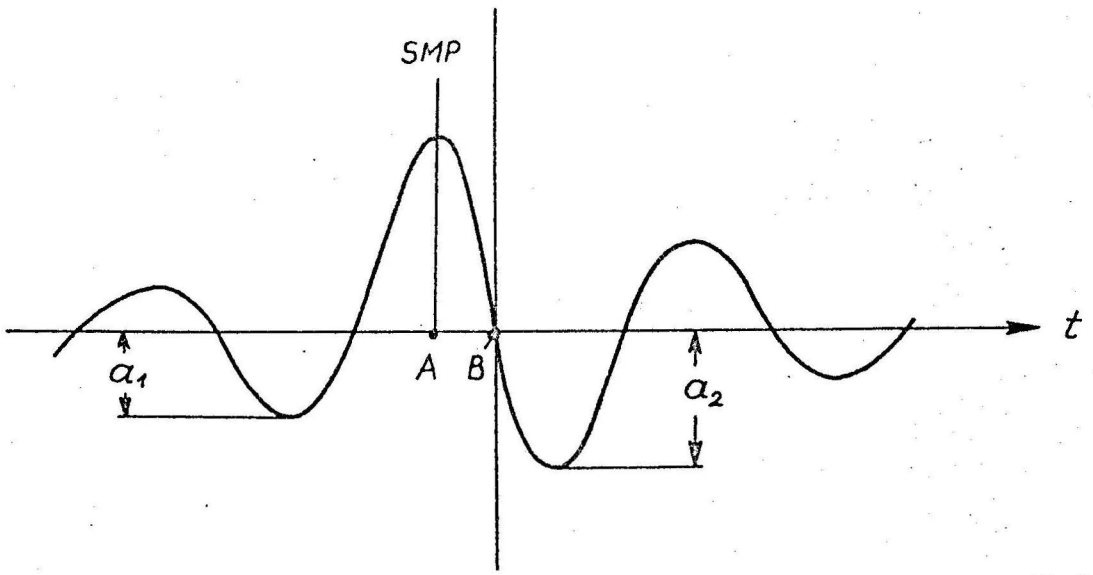


Abb.14, Nullpunktslage fuer die Entwicklung nach ungeraden Gaussfunktionen

Die Praxis hat gezeigt, dass die Approximation besser ist, wenn die Gaussfunktionen rascher abklingen. Das laesst sich durch die Wahl eines guenstigen Abszissenmassstabes erreichen: Die Gaussfunktion niedrigster Ordnung mit zwei Nullstellen ist die zweiter Ordnung. Es wurde deshalb ein Massstabsfaktor fuer das Argument der Gaussfunktionen so bestimmt, dass der mittlere Nullpunktsabstand in dem betrachteten Analysebereich mit dem Nullpunktsabstand der Gaussfunktion zweiter Ordnung uebereinstimmt.

Nachdem Nullpunkt und Abszissenmassstab festgelegt worden sind, kann die Entwicklung nach Funktionen gerader bzw. ungerader Ordnungszahl durchgefuehrt werden.

Als Ergebnis des ersten Analysedurchgangs wird nur die Gaussfunktion genommen, fuer die der dem Betrage nach groesste Koeffizient berechnet wurde.

Die Berechnung des Koeffizienten A_i erfolgt nach Gl.(23).

$$A_i = \frac{\int_{t_1}^{t_2} f(t)\, G_i(t)\, dt}{\int_{t_1}^{t_2} [G_i(t)]^2\, dt} \qquad (23)$$

mit $$f(t) = g[k(t-t_i)]$$

Dabei bedeuten g die zu approximierende Zeitfunktion,
G_i Gaussfunktionen i-ter Ordnung
k den Massstabsfaktor fuer die Argumente und
t_i die Lage des Nullpunktes.

Weil die Funktionswerte ausserhalb des betrachteten Intervalls als Null angenommen wurden, braucht die Integration nur ueber die Intervallgrenzen t_1 bis t_2 durchgefuehrt zu

werden.

Nach der Bestimmung von A_j wird von der Funktion f die mit A_j multiplizierte Gaussfunktion G_j subtrahiert.

Die gesamte Berechnung wird insgesamt dreimal durchgefuehrt, um die Zeitfunktion innerhalb des gewaehlten Segmentes durch drei Gaussfunktionen approximieren zu koennen.

Wie vom Verfasser in /10/ S.381 gezeigt wurde, konvergiert das Verfahren mit Sicherheit, sofern die $A_j \neq 0$ sind.

Codierung

Bei der Sprachsynthese aus Gaussfunktionen wird ein Parametersatz zur Erzeugung eines Pitchintervalls benoetigt.

Der Parametersatz besteht aus der Laenge der Pitchperiode, einem Zeitmassstabsfaktor und fuer die Erzeugung der drei Gaussfunktionen aus drei Nullpunktslagen, drei Amplitudenwerten und drei Ordnungszahlen.

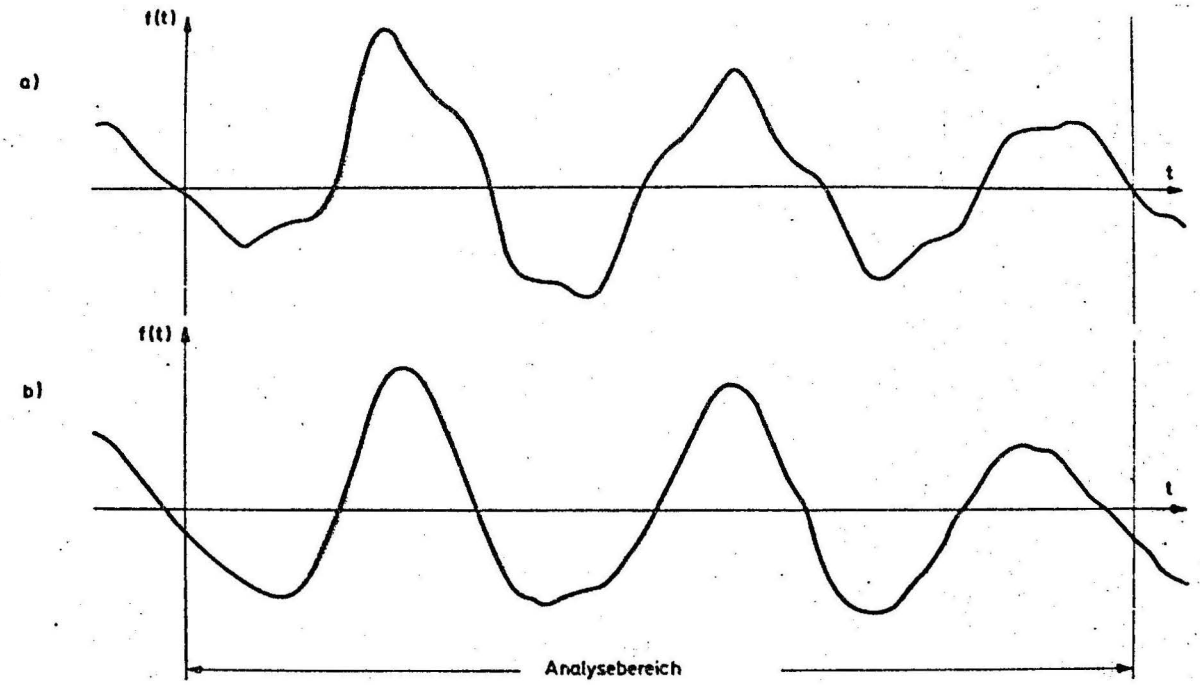
Tabelle 1 gibt die zur Codierung einer Pitchperiode benoetigte Bitzahl an.

Laenge der Pitchperiode --------	7 bit
Zeitmassstabsfaktor ----------	7 bit
3 Nullpunktslagen a 7 bit ------	21 bit
3 Amplitudenwerte a 6 bit ------	18 bit
3 Ordnungszahlen a 3 bit ------	9 bit
	62 bit

Tabelle 1, Codierung zur Sprachsynthese aus Gaussfunktionen.

Nimmt man als durchschnittliche Laenge einer Pitchperiode 10 ms an, so ergibt sich eine Uebertragungsrate von 620 b/s bzw. ein Sprachkompressionsfaktor von 129.

Die Sprachanalyse muss mit Hilfe eines Programms auf dem Digitalrechner ausgefuehrt werden.
Die Sprachsynthese erfolgte fuer ein Sprachbeispiel mit grossem Rechenzeitaufwand ebenfalls auf dem Digitalrechner. Die Abb.15a zeigt einen Ausschnitt aus einer analysierten Sprachzeitfunktion und Abb.15b zeigt die Synthese fuer diesen Fall.



a) Originalverlauf der Sprach-Zeitfunktion.
b) Approximation durch drei Gaußfunktionen innerhalb eines Analysebereiches.

Abb.15, Approximation einer Sprachzeitfunktion durch drei Gaussfunktionen.

Synthese auf dem Hybridrechner

Es wurde vom Verfasser versucht, mit Hilfe der Hybridrechenanlage CAE 90-40/ TFK, RA 770 die Gaussfunktionen aus analogen Rechenschaltungen zu gewinnen, um dadurch eine Zeitersparnis bei der Synthese herbeizufuehren (/14/,/15/)

Abb.16 zeigt den Aufbau der Rechenschaltung auf dem Analog- und dem Digitalprogrammierfeld des Rechners RA 770.

Die Abbildung zeigt in ihrem oberen Teil die Rechenschaltungen, die zur Erzeugung dreier, voneinander unabhaengiger Gaussfunktionen benoetigt werden. Die erforderlichen Amplituden werden durch die Potentiometer 46, 45, 40 und die Ordnungszahlen durch die Potentiometer 56, 66, 36 eingestellt.

Die sechs Flipflops, die in der unteren Haelfte des Bildes zu sehen sind, werden ueber Controllines vom Programm gesetzt. Sie steuern die Vorzeichen der Anfangswerte der Integrierer, sowie die Vorzeichen der resultierenden Gaussfunktionen.

Sowohl die Gaussfunktionen, als auch ihre Ableitungen fallen zu positiven und negativen Argumenten stark ab. Deshalb ist es nicht in jedem Falle moeglich, zu Beginn eines Sprachsegmentes saemtliche Anfangswerte fuer alle drei Differentialgleichungen mit hinreichender Genauigkeit einzustellen. Die Anfangswerte werden deshalb fuer feste Argumente berechnet und die Integrierer zu verschiedenen Zeitpunk-

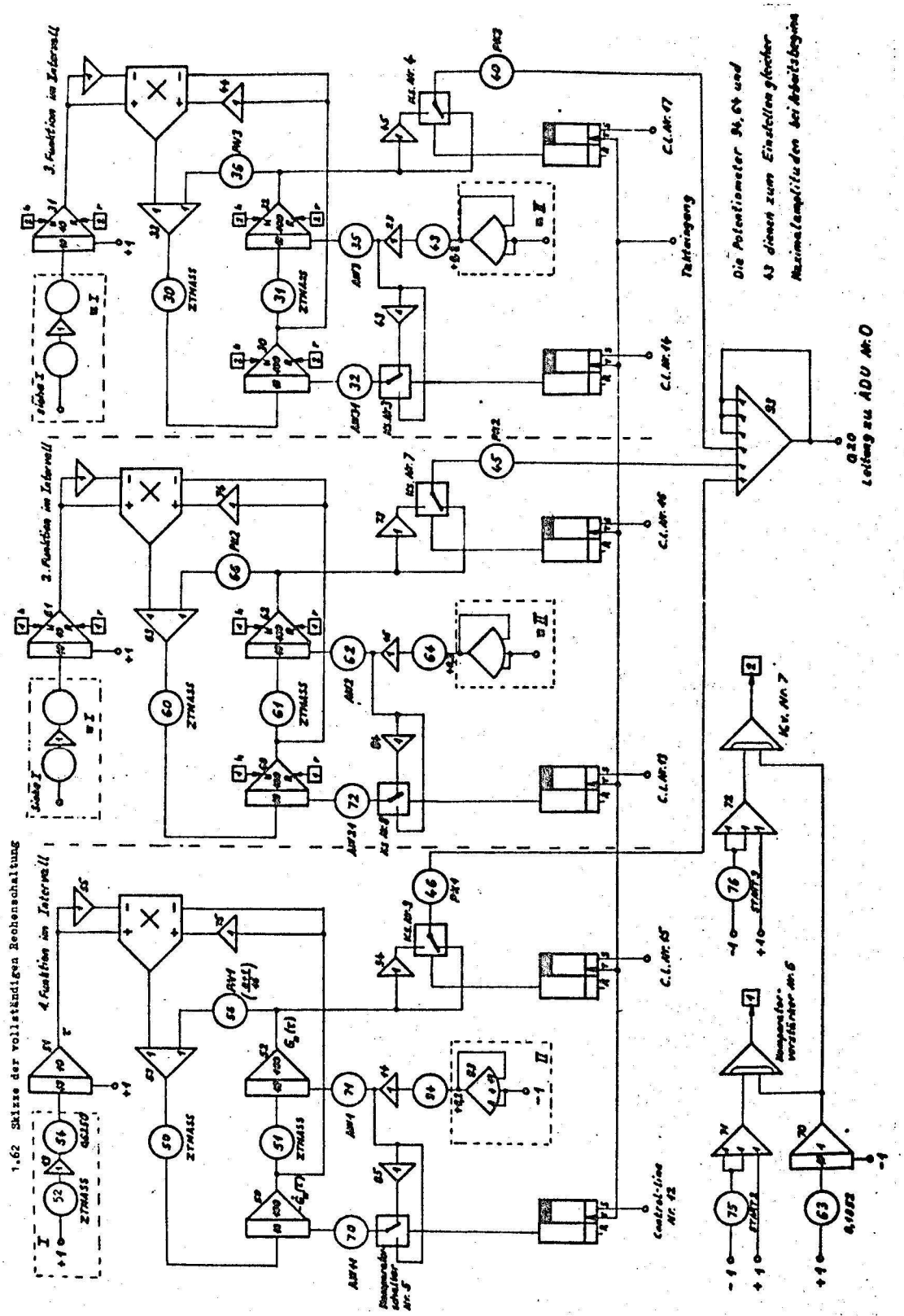


Abb.16 Spracherzeugung aus Gaußfunktionen

ten gestartet.

Die Lage der Startpunkte zueinander wird mit den Komparatoren 6 und 7 durch Vergleich einer zeitproportionalen Spannung mit fest eingestellten Spannungen definiert. Die zeitproportionale Spannung wird dabei mit dem Integrierer 70 erzeugt.

Die drei Gaussfunktionen werden ueber den Summierer 93 zur synthetischen Sprachzeitfunktion zusammengesetzt.

Die synthetische Sprache kann nur in Segmenten erzeugt werden. Fuer die kontinuierliche Spracherzeugung waere es notwendig, die Schaltung zweimal aufzubauen und benachbarte Segmente alternierend mit der einen und dann mit der anderen Schaltung zu erzeugen.

Eine weitere Moeglichkeit, Funktionsverlaeufe nach anderen Funktionen zu entwickeln, ist dann gegeben, wenn es sich um ein orthogonales Funktionssystem handelt. Vom mathematischen Standpunkt ist die Entwicklung nach orthogonalen Funktionen uebersichtlicher und eleganter, jedoch ist man dann an bestimmte Funktionsverlaeufe gebunden. Ausserdem muss unter Umstaenden die zu approximierende Funktion aus einer groesseren Anzahl Teilfunktionen zusammengesetzt werden, als es bei dem oben beschriebenen Verfahren der Fall ist.

3.2 Autokorrelationsvocoder

Wenn x(t) ein stationaeres Rauschsignal darstellt, ist die Autokorrelationsfunktion definiert durch die Gleichung:

$$\varphi(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)\cdot x(t+\tau)\,dt \quad (24)$$

Nach Wiener steht die Autokorrelationsfunktion durch

$$\varphi(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi(\omega)\, e^{j\omega\tau}\, d\omega \quad (25)$$

in fester Beziehung zum Leistungsdichtespektrum . Filtert man die Zeitfunktion q(t) mit dem Spektrum

$$Q(s) = \mathcal{L}[q(t)] \quad (26)$$

mit einem Filter der Uebertragungsfunktion H(s) und der Impulsanwort

$$h(t) = \mathcal{L}^{-1}[H(s)] \quad (27)$$

ergibt sich das Leistungsdichtespektrum:

$$\phi(s) = k_1 |Q(s)|^2 \cdot |H(s)|^2 \quad (28)$$

und die Autokorrelationsfunktion

$$\varphi(\tau) = k_2\, \mathcal{L}^{-1}[\phi(s)] \quad (29)$$

Dabei stellen k_1 und k_2 Konstanten dar.

Fuer den Fall Q(s)=1, d.h. fuer den Fall, dass die Quelle weisses Rauschen abgibt, stellt die Autokorrelationsfunktion die Impulsantwort des Filters dar, wobei der Spektralverlauf des Filters quadriert worden ist. Die urspruenglichen Phasenbeziehungen des Filters nach Gl.(27) sind durch die Quadrierung im Spektrum nicht mehr vorhanden.

Bei den Gleichungen zur Beschreibung des menschlichen Spracherzeugungssystems nach Gl.(18) und Gl.(19) moege die Abstrahlung den Filtereigenschaften des Vokaltraktes zugeschlagen werden. Dann vereinfacht sich die Gl.(19) zu:

$$P(s) = Q(s)\cdot H(s) \quad (30)$$

und

$$p(t) = \mathcal{L}^{-1}[P(s)] \quad (31)$$

Aus der Gegenueberstellung von Gl.(28), Gl.(29) und Gl.(30), Gl.(31) ergibt sich, dass die Autokorrelationsanalyse von Sprache die Impulsantwort eines Filters liefert, das in seinen Uebertragungseigenschaften groesste Aehnlichkeit mit dem Vokaltrakt aufweist. Das gilt natuerlich nur unter der Bedingung, dass der Einfluss der Quelle als Q(s) in Gl.(28)

vernachlaessigt werden kann.

Charakteristika in der Uebertragungsfunktion des Vokaltraktes sind die ausgepraegten Maxima, die bei den Frequenzwerten der Formanten auftreten. Wenn die Uebertragungsfunktion quadriert wird, bleiben die ausgepraegten Maxima, d.h. die Formanten, in ihrer richtigen Frequenzlage erhalten. Aus diesem Grunde aehnelt das Spektrum der Autokorrelationsfunktion der Uebertragungsfunktion des Vokaltraktes.

Prinzip des Autokorrelationsvocoders

Auf der Aehnlichkeit zwischen dem Spektrum der Autokorrelationsfunktion und der Uebertragungsfunktion des Vokaltraktes basiert die Funktionsweise des Autokorrelationsvocoders.

In Anlehnung an Gl.(18) und Gl.(19) mit der Modifikation nach Gl.(30) wird das menschliche Spracherzeugungssystem, bestehend aus einem Generator und einem nachgeschalteten Sprachfilter simuliert.

Der Generator gibt im Falle stimmhafter Laute eine Pulsfolge und im Falle stimmloser Laute Rauschen ab.

Dem Generator ist ein Filter nachgeschaltet, das in jedem Augenblick dem Vokaltrakt aehnliche Uebertragungseigenschaften aufweisen soll. Die Uebertragungseigenschaften des Filters werden durch die Impulsantwort des Filters charakterisiert. Die diskreten Werte der Impulsantwort sind in diesem Falle die diskreten Werte der Kurzzeitautokorrelationsfunktion, die aus dem Sprachsignal ermittelt werden.

Sprachsynthese

Die Uebertragungsfunktion eines digitalen Filters mit dem Abtastintervall T lautet allgemein:

$$H(z) = \frac{\sum_{\mu=0}^{m} b_\mu z^\mu}{\sum_{\nu=0}^{n} c_\nu z^\nu} \qquad z = e^{sT} \qquad (32)$$

Unter der Bedingung

$$c_n = 1 \text{ und } c_\nu = 0 \text{ für } \nu = 0, 1, \ldots, (n-1)$$

erhaelt man

$$H^*(z) = \frac{1}{z^n} \sum_{\nu=0}^{n} b_\nu \cdot z^\nu \qquad (33)$$

Die Abb.17 zeigt die der zweiten kanonischen Form entsprechende Blockschaltung. Ein derartig aufgebautes nichtrekursives digitales Filter nennt man auch Transversalfilter.

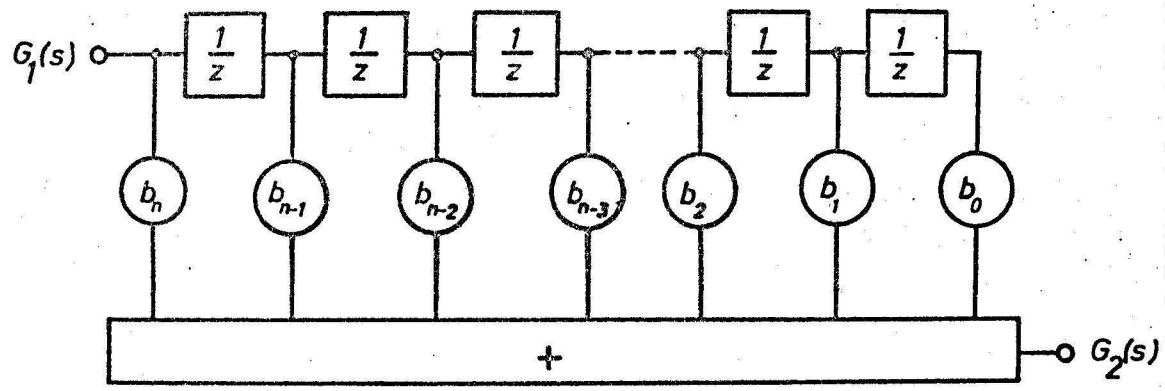


Abb.17, Transversalfilter

Die Glieder b_i sind Bewertungskoeffizienten, deren Groesse den diskreten Werten der Impulsantwort des Transversalfilters entspricht. Da beim Autokorrelationsvocoder die Impulsantwort des Sprachfilters gerade in Form diskreter Werte vorliegt, empfiehlt es sich, zur Synthese ein derartiges Transversalfilter zu verwenden und die Bewertungskoeffizienten in einer Hardwareschaltung beispielsweise durch elektronische Potentiometer zu realisieren.

Die Glieder 1/z sind Verzoegerungsglieder, die diskrete Funktionswerte um die Zeit T verzoegern. Diese Verzoegerungselemente werden auch Abtast- Halteglieder genannt. Eine hardwaremaessige Realisierung findet sich in /18/.

Die Taktfrequenz fuer die Abtast- Halteglieder bestimmt die hoechste Frequenz und die Durchlaufzeit durch saemtliche Glieder die niedrigste Filterfrequenz des Transversalfilters. Die Anzahl der Abtast- Halteglieder ist damit gleich dem Quotient der hoechsten zur niedrigsten uebertragenen Frequenz. Geht man von Telephonqualitaet des Sprachfilters aus, d.h. einem Frequenzbereich von 300 - 3000 Hz, so sind 10 Abtast- Halteglieder erforderlich.

Der Autokorrelationsvocoder nach CHRISTIANSEN und SCHWEIZER /19/ benoetigt im Gegensatz dazu 24 Verzoegerungseinheiten.

Sprachanalyse

Die Parameter, die im Analyseteil des Vocoders ermittelt werden und im Syntheseteil die kuenstliche Spracherzeugung steuern, sind:

1. Stimmhaft- Stimmlosigkeit
2. Pitchfrequenz
3. .
. } diskrete Werte der Autokorrelationsfunktion
. .
n

Die Bestimmung der Parameter 1. und 2., die unter dem Namen 'Pitchbestimmung' zusammengefasst ist, wird im Kap.6 noch sehr ausfuehrlich behandelt. Deshalb soll an dieser Stelle nur auf die Bestimmung der diskreten Werte der Autokorrelationsfunktion eingegangen werden.

Nach FLANAGAN (/2/ S.133) kann man die Kurzzeitautokorrelationsfunktion schreiben als:

$$\varphi(\tau,t)=\int_{-\infty}^{t} f(\lambda)\cdot f(\lambda+\tau)\cdot k(t-\lambda)\,d\lambda \qquad (34)$$

dabei ist k(t) eine Gewichtsfunktion mit der Bedingung

$$k(t)=0 \text{ fuer } t<0.$$

Die Abb.18 zeigt in einem Blockschaltbild eine Moeglichkeit, wie man einen einzelnen Wert $\varphi(\tau,t)$ der Kurzzeitautokorrelationsfunktion ermitteln kann.

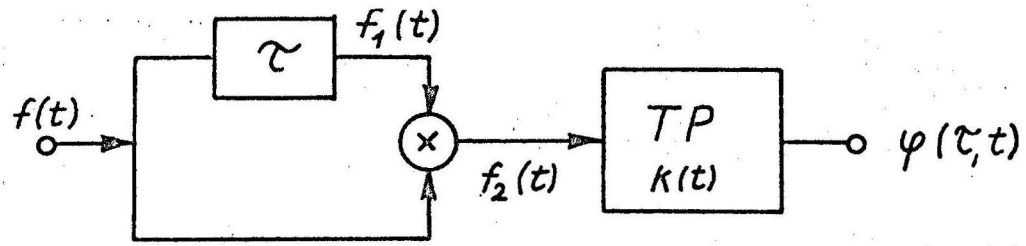


Abb.18, Bestimmung der Kurzzeit-autokorrelationsfunktion

Am Ausgang des Verzoegerungsgliedes liegt das Signal $f_1(t)$ vor, das gegenueber f(t) um τ verzoegert ist. Der Multiplizierer erzeugt das Signal $f_2(t)=f_1(t)\cdot f(t)$.
Die Mittelung von $f_2(t)$ ueber eine feste Zeit, die zugleich auch einer Bewertung mit dem Zeitfenster k(t) entspricht, erfolgt durch den Tiefpass TP.

Ersetzt man das Verzoegerungsglied durch eine Verzoegerungsleitung mit mehreren Abgriffen und fuehrt so viele Multiplikatoren und Tiefpaesse ein, wie diskrete Werte der Kurzzeitautokorrelationsfunktion gewuenscht werden, ergibt sich die Analyseschaltung nach Abb.19.

Die Analyseschaltung nach Abb.19 kann hardwaremaessig aufgebaut werden und in Realzeit arbeiten.

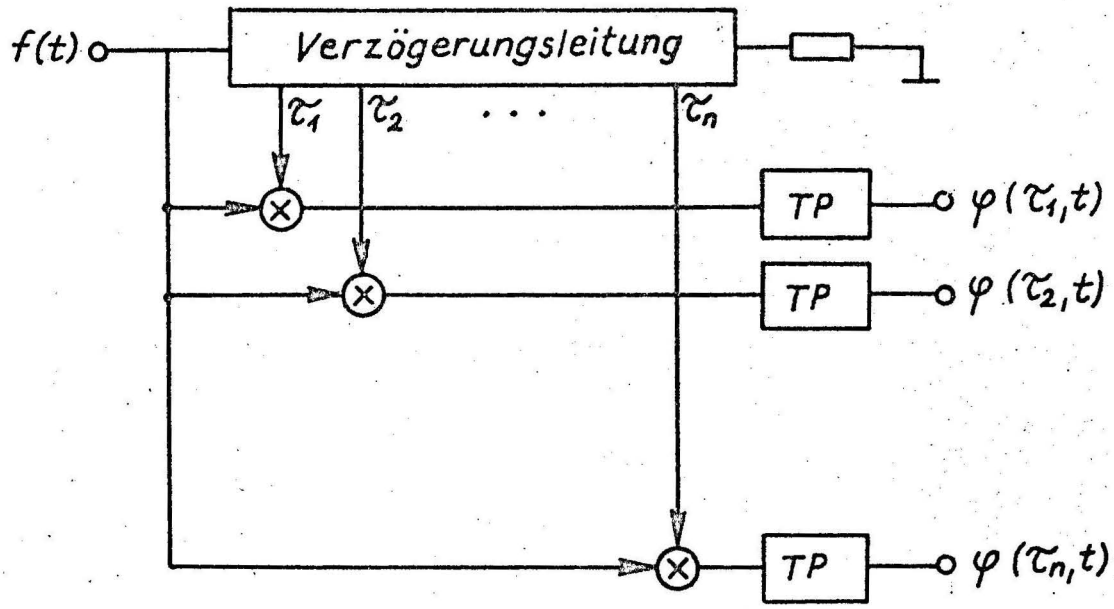


Abb.19, Ermittlung diskreter Werte der AKF

Soll die Analyse durch ein Programm auf dem Digitalrechner durchgefuehrt werden, ergibt sich noch eine einfachere Moeglichkeit zur Bestimmung der Kurzzeitautokorrelationsfunktion:
Die Sprachzeitfunktion, die in Form von Abtastwerten vor-

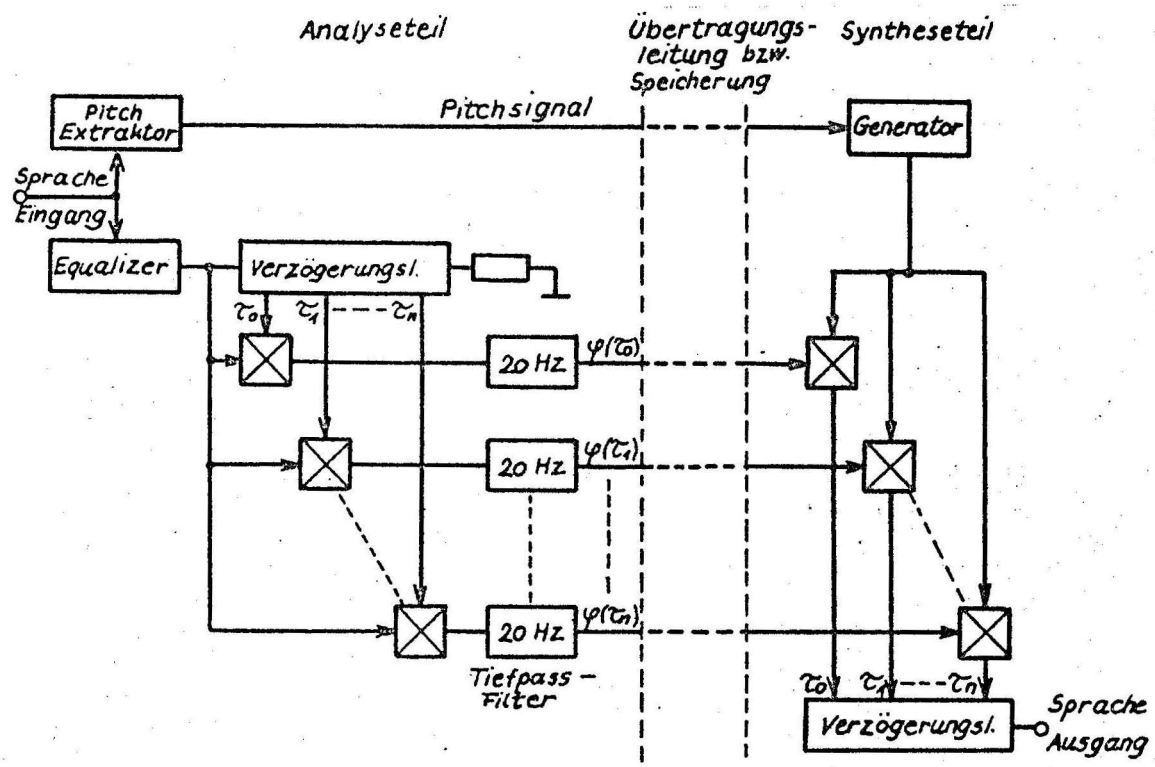


Abb.20, Autokorrelationsvocoder

liegt, wird mit einem Zeitfenster bewertet, das den Zeit-

punkt charakterisiert, fuer den gerade die Autokorrelationsfunktion berechnet werden soll. Um den Aliasing- Effekt zu verhindern, hat das Zeitfenster beispielsweise die Form eines Hamming- Windows (/16/ S.296). Die Kurzzeitautokorrelationsfunktion berechnet sich dann zu:

$$\varphi(\tau;t) = k \cdot \mathcal{L}^{-1}\{|\mathcal{L}[h(\tau;t)\cdot f(t)]|^2\} \quad (35)$$

Fuer den Fall diskreter Zeitwerte kann die Hin- und Ruecktransformation, die in Gl.(35) fuer den kontinuierlichen Fall durch die Laplacetransformation angedeutet wurde, mit dem FFT- Algorithmus nach COOLEY - TUKEY /17/ ausgefuehrt werden.

Die Abb.20 zeigt die vollstaendige Darstellung eines Autokorrelationsvocoders.

Tabelle 2 gibt die vom Verfasser geschaetzte Bitzahl an, die mindestens zur Codierung der Steuerparameter benoetigt wird.

Stimmhaft- Stimmlosigkeit --------	2 bit/20 ms
Pitchfrequenz ----------	4 bit/20 ms
Ausgangsamplitudenwert ----------	6 bit/20 ms
10 Amplitudenwerte der Autokorrelationsfunktion a 6 bit -	60 bit/20 ms
	72 bit/20 ms

Tabelle 2, Codierung des AKF- Vocoders

Das entspricht einer Uebertragungsrate von 3600 bit/sek und einem Kompressionsfaktor von 22.

Sprachentzerrung

Lt. SCHROEDER /7/ und WINCKEL /8/ hoert sich die mit einem Autokorrelationsvocoder synthetisierte Sprache 'knallig' verzerrt an. Die Ursache liegt darin, dass das Spektrum dem Quadrat des Spektrums des menschlichen Vokaltraktes entspricht. Dadurch erscheinen zwar stark ausgepraegte Formanten sehr betont, schwaechere Formanten verschwinden aber vollstaendig.

Der Effekt kann dadurch beseitigt werden, indem die Sprache zunaechst durch einen Wurzelzieher in dem Sinne vorverzerrt wird, dass der Spektralverlauf der Quadratwurzel des Spektrums des Vokaltraktes entspricht. In dem Falle wird die Sprache durch den nachgeschalteten Autokorrelationsvocoder wieder entzerrt.
Ein Blockschaltbild fuer einen derartigen Wurzelzieher nach SCHROEDER (/7/ S.728) zeigt die Abb.21.

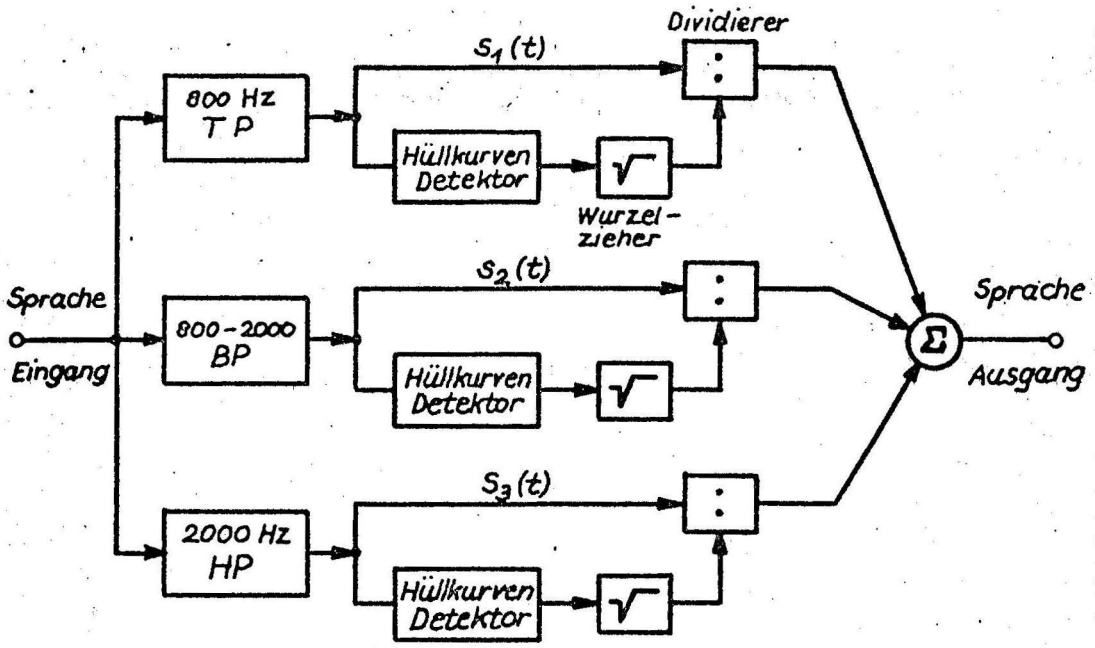


Abb.21, Wurzelzieher /7/

3.3 Kanalvocoder

Der Kanalvocoder ist der bekannteste und verbreitetste Vocodertyp (/2/,/22/,/25/). Er baut in seiner Funktionsweise auf Gl.(30) auf.

Die Quelle entspricht der, die unter 3.2 beschrieben wurde.

Zur Simulation des Vokaltraktes wird ein Filter benoetigt, das in jedem Augenblick die Uebertragungseigenschaften des Vokaltraktes aufweist. Die Phaseninformation der Uebertragungsfunktion wird dabei vernachlaessigt. Das ist zulaessig, da Versuche die relativ geringe Bedeutung der Phaseninformation fuer die Verstaendlichkeit von Sprache nachgewiesen haben. Ein typisches Beispiel fuer den Frequenzgang des Vokaltraktes stellt G_{dB1} in Abb.22 dar.

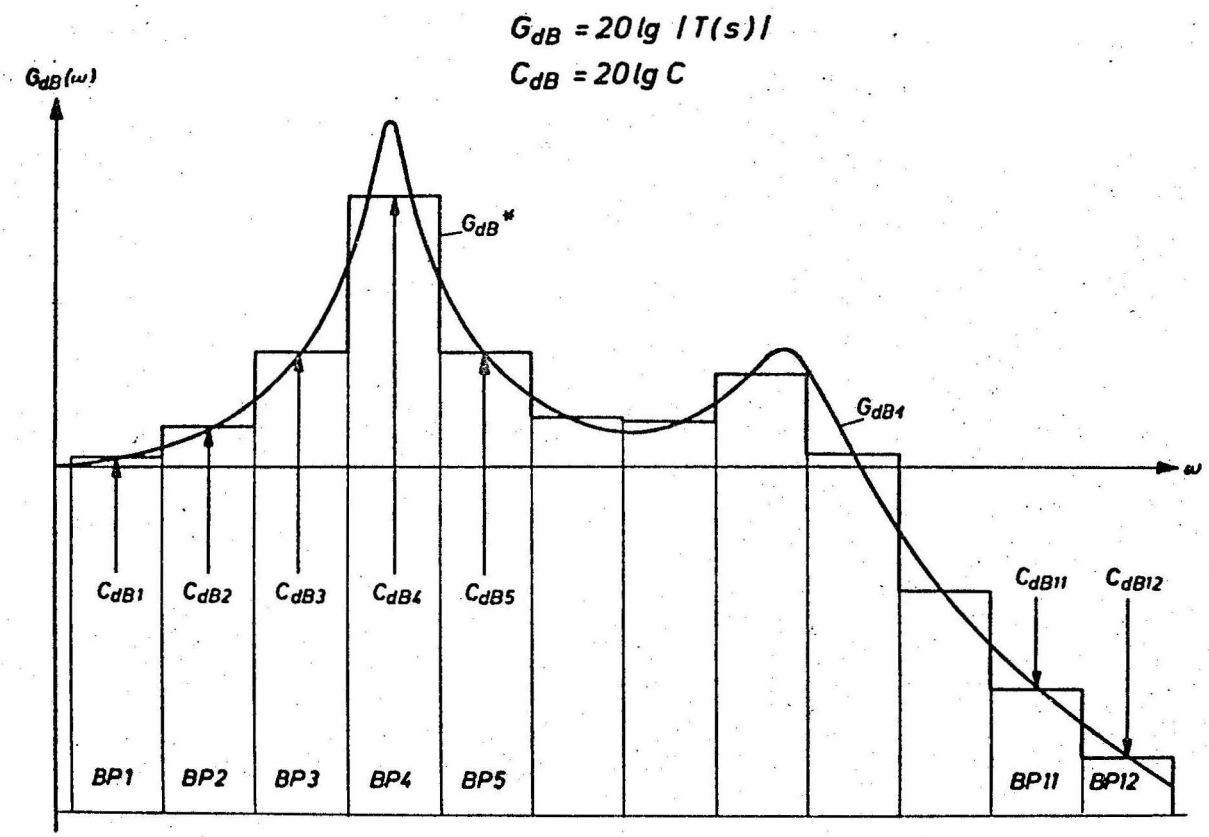


Abb.22, Approximation des Betrages der Uebertragungsfunktion des Vokaltraktes durch eine Treppenkurve

Synthese

Beim Kanalvocoder wird der Frequenzbereich des Vokaltraktes durch eine sog. Filterbank in n Abschnitte unterteilt. Die

Filterbank besteht aus n parallelgeschalteten Bandpaessen. Die oberen und unteren Grenzen der benachbarten Bandpaesse werden gerade so gewaehlt, dass sie sich in ihrem 3-dB-Abfall ueberschneiden.

Abb.23 zeigt das Blockschaltbild eines Kanalvocoders. Wird auf die Eingaenge saemtlicher Bandpaesse (in Abb.23 rechts vom Uebertragungskanal) ein Puls gegeben, so schwin-

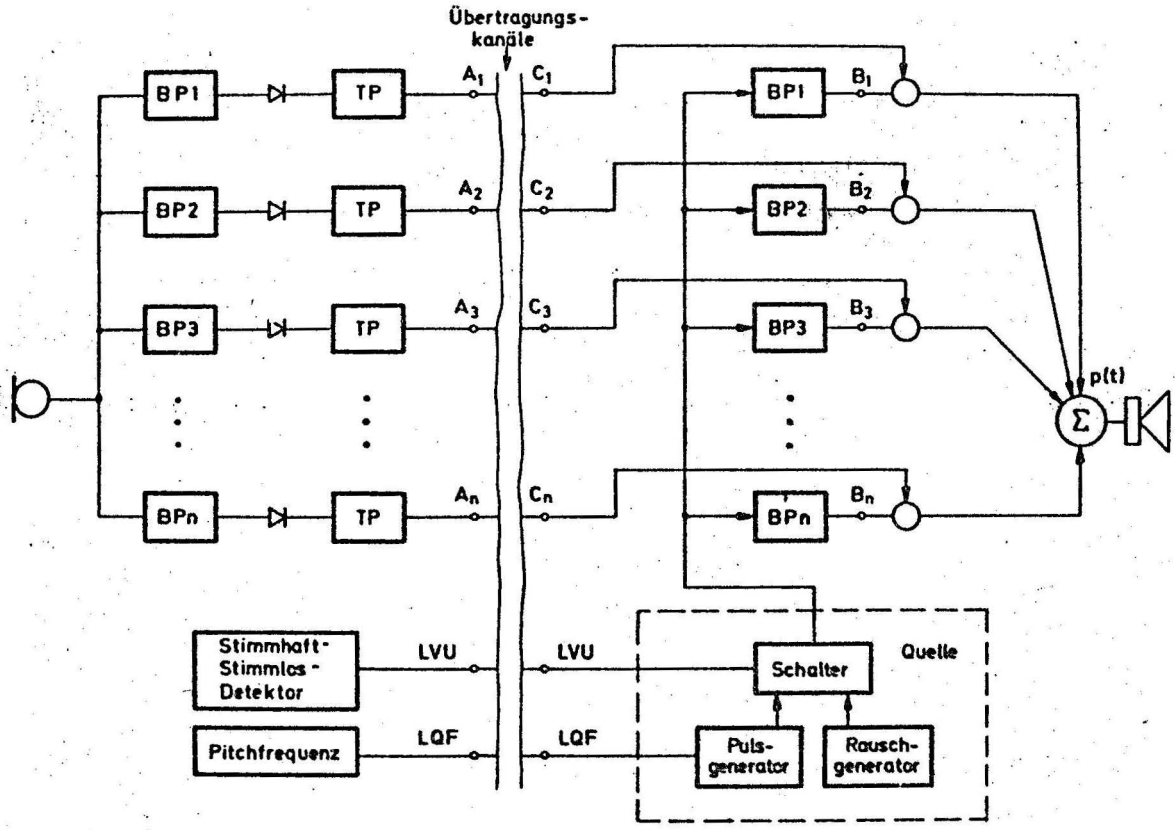


Abb.23, Kanalvocoder

gen alle Bandpaesse mit gleicher Amplitude aber unterschiedlicher Frequenz, denn jeder Bandpass schwingt mit seiner Mittenfrequenz. Wuerde man die Ausgangssignale B_1 bis B_n aufsummieren, erhielte man eine Zeitfunktion, die saemtliche (Mitten-)Frequenzen gleichmaessig uebertraegt.

Bewertet man die Ausgaenge der Bandpaesse z.B. durch Koeffizienten C_i, deren Groesse man aus Abb.22 im logarithmischen Massstab als $C_{dB;}$ entnehmen kann, dann werden im Summensignal p(t) nicht mehr alle Frequenzen gleich stark vertreten sein, sondern die einzelnen Frequenzen werden je nach Bewertung durch die C_i erscheinen. Damit uebertraegt die Filterbank nicht mehr alle Frequenzen gleichmaessig, sondern sie stellt ein Filter dar, dessen Frequenzgang sich durch die Koeffizienten C_i als Treppenkurve G_{dB}^* (s. Abb.22) einstellen laesst. So kann man jeden Frequenzgang approximieren.

Die Approximation wird dabei um so besser sein, je mehr

Bandpaesse vorliegen, d.h. auch je schmalbandiger die Filter werden. Je mehr Bandpaesse verwendet werden, um so mehr Koeffizienten C_i muessen uebertragen werden und desto geringer wird der Sprachkompressionsfaktor.

Im Zusammenhang mit der vorliegenden Arbeit wurde ein Kanalvocoder auf dem Digitalrechner simuliert, der aus 15 Kanaelen besteht /20/.

Bei der Simulation einer Filterbank benutzt man vorzugsweise die 'Frequency- Sampling- Technique'(/21/ S.81).

Ausgangspunkt fuer die Frequency- Sampling- Technique ist ein sog. Kammfilter. Es hat die z-Uebertragungsfunktion:

$$H(z) = 1 + z^{-m} \quad (36)$$

Das Kammfilter weist laengs des Einheitskreises in konstantem Abstand Nullstellen auf. Die Lage der Nullstellen ist:

$$z_k = \exp\left[\frac{j2\pi(K+1/2)}{m}\right] \qquad k = 0,1,\ldots,m-1 \quad (37)$$

Schaltet man in Reihe mit einem Kammfilter einen einfachen Resonator, der ein konjugiert komplexes Polpaar enthaelt und dessen Polpaar gerade ein Nullstellenpaar des Kammfilters kompensiert, gewinnt man einen einfachen Bandpass. Die Flan-

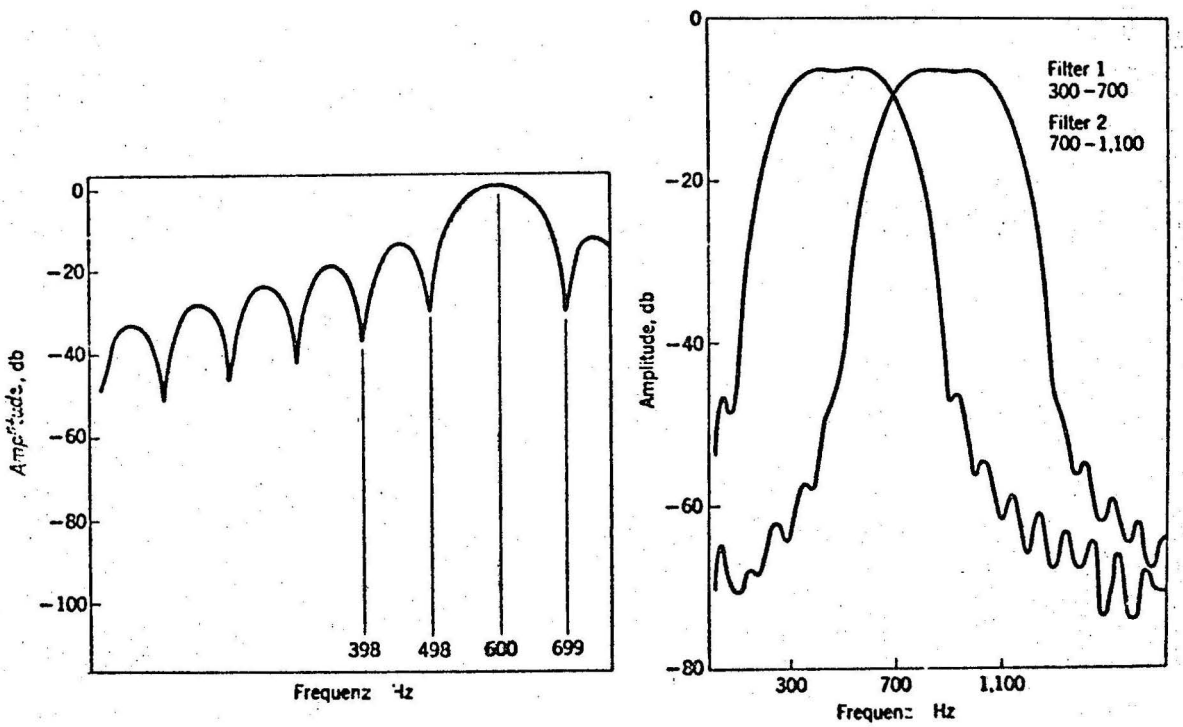


Abb.24, Frequenzgang eines Kammfilters in Reihe mit 1 bzw. 7 Resonatoren /21/

kenstellheit und der Betrag der Uebertragungsfunktion im Durchlassbereich lassen sich durch geeignete Kombination mehrerer parallel geschalteter Resonatoren erheblich verbessern. Die Abb.24a zeigt den Frequenzgang des Kammfilters in Reihe mit einem einzelnen Resonator und Abb.24b in Reihe mit sieben kombinierten Resonatoren. Durch Zusammenfassung mehrerer Resonatorkombinationen kann eine Filterbank aus benachbarten Bandpaessen aufgebaut werden.
Das Blockschaltbild nach Abb.25 zeigt, wie jeweils drei benachbarte Resonatoren, die hier mit P_i abgekuerzt sind, ueber jeweils ein Summierglied zu einem Bandpass zusammengefasst werden koennen (/5/ S.1362).

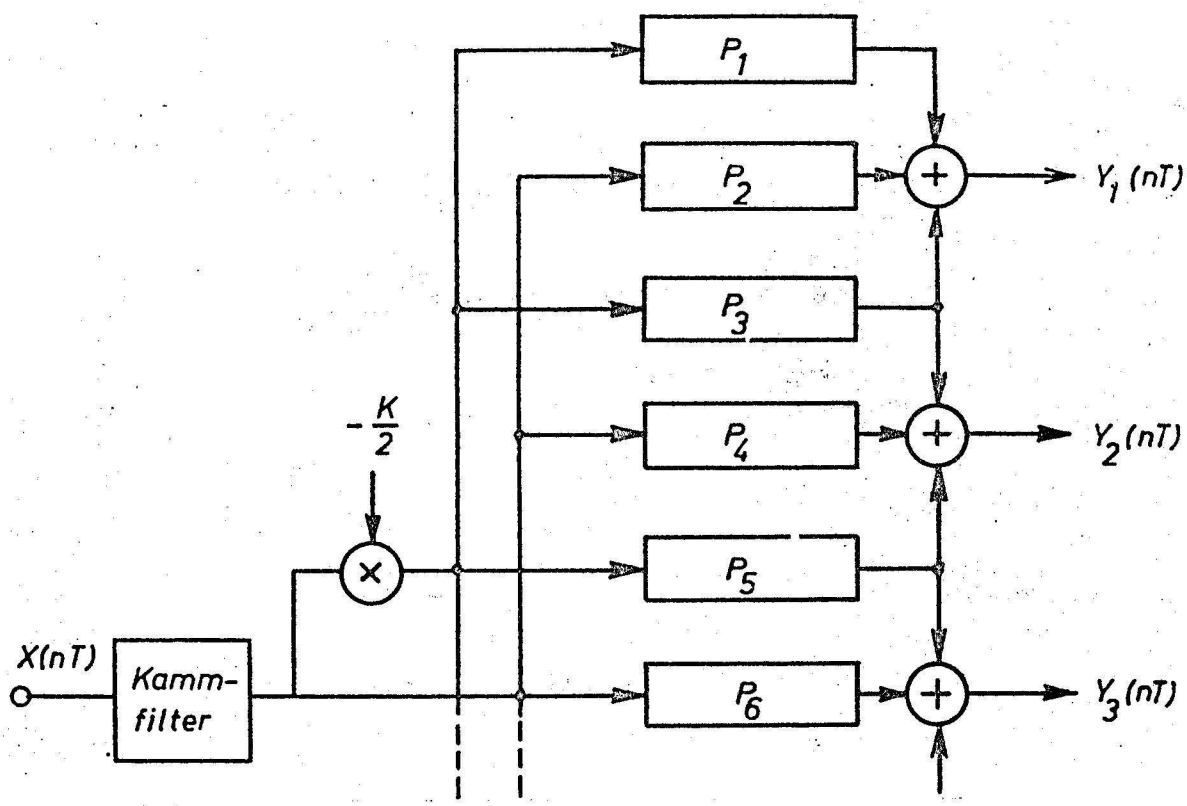


Abb.25, Aufbau einer Filterbank in Frequency-Sampling- Technique

Der Nachteil einer solchen Filterbank liegt darin, dass benachbarte Bandpaesse in ihrer Bandbreite nicht beliebig variiert werden koennen. Es hat sich naemlich gezeigt, dass unter Beruecksichtigung der Hoereigenschaften des menschlichen Ohres in den niedrigen Frequenzbaendern mehr Information enthalten ist, als in den hoeheren. Aus diesem Grunde wurde nach SCHWEIZER /22/ ein Kanalvocoder aufgebaut, bei dem die Bandbreiten der Kanaele gleich den Frequenzgruppen (nach ZWICKER) gemacht wurden. Er benoetigte dabei fuer den Frequenzbereich von 300 - 3400 Hz 14 Kanaele.

Analyse

Mit der oben beschriebenen Syntheseschaltung laesst sich ein Filter aufbauen, das in jedem Augenblick den Frequenzgang des Vokaltraktes nachbilden kann unter der Voraussetzung, dass auch fuer jeden Augenblick die Parameterkombination C_i bekannt ist. Die Analyseschaltung ist in Abb.23 links von den Uebertragungskanaelen zu sehen. Die Sprachzeitfunktion, die z.B. durch Abtastung einer Mikrophonspannung gewonnen werden kann, wird auf eine Filterbank gegeben. Die Filterbank ist genauso aufgebaut wie die, die bereits in der Syntheseschaltung beschrieben wurde. Die Ausgaenge der Bandpaesse werden gleichgerichtet und durch Tiefpaesse geglaettet. Die Ausgangssignale A_i sind dann zu jedem Augenblick dem Betrag des Spektrums im Durchlassbereich des entsprechenden Bandpasses proportional. Da die Bandpaesse aber den ganzen Frequenzbereich lueckenlos ueberstreichen, stellen die Ausgaenge A_i die Approximation des Spektrums durch eine Treppenkurve dar, wie es in Abb.22 als G^{*}_{dB} dargestellt wurde. Daraus ergibt sich, dass die A_i gerade die Koeffizienten sind, die als Parameter c_i zur Steuerung des Syntheseteils benoetigt werden.

Simuliert man einen Kanalvocoder auf dem Digitalrechner, so kann man die gesuchten Steuerparameter fuer die Synthese recht einfach dadurch erhalten, indem man aus der Zeitfunktion mit Hilfe einer Fouriertransformation das Spektrum berechnet. Wie Abb.22 zeigt, wird das Spektrum in die einzelnen Kanalabschnitte zerlegt und durch Mittelwertbildung innerhalb eines Kanalabschnittes die gesuchte Huellkurvenamplitude ermittelt.
Die Steuerparametersaetze werden bei der Realzeitanalyse als Ausgangsamplituden einer Filterbank zu aequidistanten Zeitpunkten abgetastet bzw. bei der Simulation fuer aequidistante Zeitpunkte berechnet und anschliessend quantisiert.

Nach SCHWEIZER genuegt es, jeden Kanal mit 3 bit zu quantisieren, wenn eine logarithmische Amplitudenstufung vorgenommen wird.

Aus der Tabelle 3 ergibt sich, dass eine Uebertragungsrate von 2400 bit/sek und damit ein Kompressionsfaktor von 33 erreicht werden kann.

Stimmhaft- Stimmlosigkeit -----	2 bit/20 ms
Pitchfrequenz ----------	4 bit/20 ms
14 Kanaele a 3 bit ----------	42 bit/20 ms
	48 bit/20 ms

Tabelle 3, Codierung des Kanalvocoders

Der Analyse- und Syntheseteil eines Kanalvocoders koennen hardwaremaessig aufgebaut werden und in Realzeit arbeiten. Der Aufbau der Filterbaenke in Hardware ist umfangreich, insbesondere, da die vorteilhafte Frequency- Sampling Technique nur bei den in Hardwarerealisierung bisher ungebraeuchlichen digitalen Filtern moeglich ist.

3.4 Formantvocoder

Der Formantvocoder hat eine starke Aehnlichkeit mit dem Kanalvocoder. Die Quelle, die zur Anregung des Sprachfilters dient, ist mit der des Kanalvocoders identisch. Waehrend der Kanalvocoder ein Sprachfilter enthaelt, das die Huellkurve des Kurzzeitspektrums der Sprache uebertraegt, beschraenkt sich der Formantvocoder darauf, charakteristische Maxima des Kurzzeitspektrums, die sog. Formanten, zu uebertragen.

Die besondere Bedeutung der Formanten geht aus Gl.(12) in 2.2 hervor. Gl.(12) besagt, dass bei der Artikulation stimmhafter Laute die Uebertragungsfunktion durch unendlich viele Formanten, aber keine Antiformanten beschrieben wird. Praktische Untersuchungen haben noch zusaetzlich ergeben, dass hoechstens die ersten fuenf Formanten bei der Spracherzeugung von Bedeutung sind und dass dabei wiederum nur die ersten drei Formanten variiert werden brauchen.

Aus 2.3 geht hervor, dass die Formantstruktur bei allen anderen Lauten ausser den Vokalen mehr oder weniger verlorengeht. Die Uebertragungseigenschaften des Sprachfilters koennen in diesem Falle durch den Einbau von Antiformanten den Eigenschaften des Vokaltraktes angepasst werden.

Sehr haeufig werden im Synthetisator zur Erzeugung der unterschiedlichen Lautkategorien verschiedene Filterzweige im Sprachfilter vorgesehen. Ein typisches Beispiel eines derartigen Formantsynthetisators nach FANT /24/ zeigt die Abb.26.

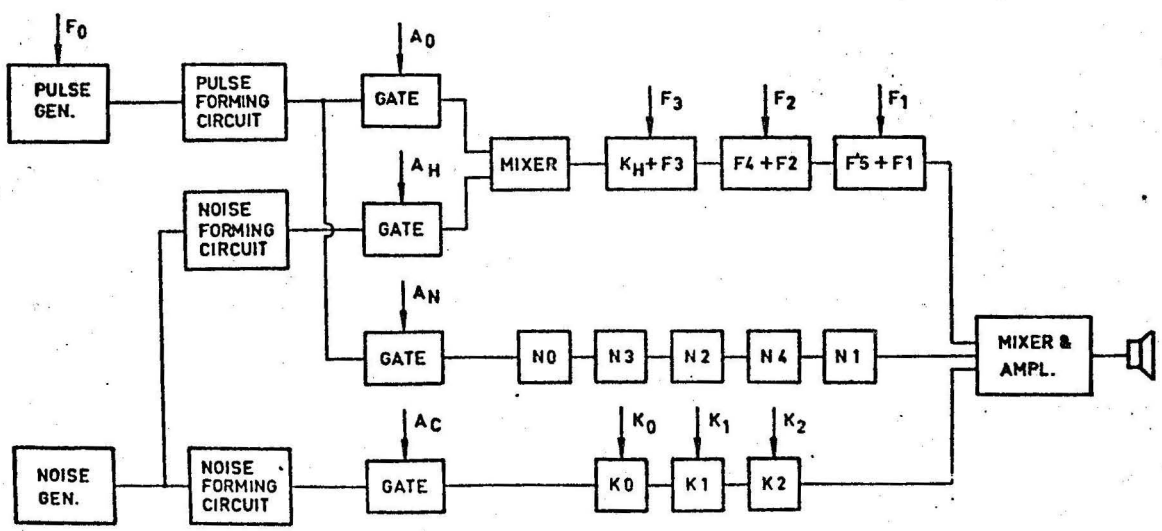


Abb.26, OVE II - Formantvocoder /24/

Die akustische Quelle zur Anregung des Sprachfilters besteht aus einem Rauschgenerator und einem Pulsgenerator. Durch eine Pulsformschaltung wird der Zeitverlauf der Pulse dem Zeitverlauf der Schnelle an der Glottis angepasst. Der Rauschgenerator wird durch die 'noise forming circuit' gefiltert und in seinem Spektralverlauf der akustischen Quelle im Vokaltrakt angeglichen.

Das eigentliche Sprachfilter ist zur Erzeugung der drei Lautkategorien: Vokale, Nasale und stimmlose Laute, in drei Zweige aufgeteilt, das sog. F- Filter zur Erzeugung von Vokalen, das N- Filter zur Erzeugung von Nasalen und das K-Filter zur Erzeugung von stimmlosen Lauten.

Das F- Filter besteht aus fuenf Formanten, von denen die ersten drei steuerbar sind. Als Anregungsquelle kann ueber die Amplitudenkontrolle A_0 entweder der Pulsgenerator oder zur Erzeugung von Fluesterlauten ueber A_H der Rauschgenerator geschaltet werden.

Das N- Filter zur Erzeugung von Nasalen besteht aus konstanten Formant- und Antiformantgliedern und kann ueber A_N vom Pulsgeneratorzweig angeregt werden.

Das K- Filter, das den Vokaltrakt waehrend der Erzeugung stimmloser Laute repraesentiert, besteht aus zwei Formanten und einem Antiformanten, die alle in ihren Frequenz- und Bandbreitewerten variiert werden koennen. Die Anregung erfolgt ueber die Amplitudenkontrolle A_C vom Rauschgenerator.

Wie aus der Abb.26 ersichtlich ist, sind zur Steuerung des Formantvocoders 11 Parameter notwendig, wobei die Steuerparameter fuer die Formant- und Antiformantfilter jeweils einen Frequenz- und einen Bandbreitewert beinhalten.

Ein in seinem Aufbau noch komplizierterer Formantsynthetisator, der ausser den beim OVE II beruecksichtigten Lautkategorien noch die Moeglichkeit bietet, zwischen stimmlosen und stimmhaften Fricativen zu unterscheiden, und bei dem Vorkehrungen zur Erzeugung stimmhafter Plosive getroffen wurden, ist von L.RABINER /23/ beschrieben worden.

Diese beiden Formantsynthetisatoren dienen vor allem fuer Untersuchungen auf dem Gebiet der kuenstlichen Spracherzeugung. Sie synthetisieren Sprache hoher Qualitaet.

Fuer die im Zusammenhang mit der vorliegenden Arbeit gestellten Anforderungen genuegt z.B. ein einfacher Formantsynthetisator nach Abb.27.

Der Aufbau des Formantvocoder, der in der Abb.27 dargestellt ist, basiert auf Gl.(18) und Gl.(19).

Der Generator besteht aus einem Rauschgenerator und einem Pulsgenerator mit nachgeschaltetem Pulsformnetzwerk. Der Pulsgenerator- oder Rauschgeneratorzweig koennen wahlweise ueber einen Schalter auf das nachfolgende Filter gegeben werden.

Das Filter enthaelt nur einen einzigen Zweig, der den Vokaltrakt bei der Erzeugung saemtlicher Lautkategorien repraesentiert. Es besteht aus fuenf Formanten, wobei die ersten drei Formanten in ihrer Frequenz variierbar sind. Wie Untersuchungen gezeigt haben, koennen die Bandbreitewerte

der Formanten aus einer Tabelle den Frequenzwerten zugeordnet werden und stellen dadurch keine unabhaengigen Parameter dar.
Wie schon oben erwaehnt wurde, wird die Uebertragungsfunktion des Vokaltraktes auch bei solchen Sprachlauten durch fuenf Formanten approximiert, die keine Formantstruktur aufweisen. Es zeigt sich, dass dadurch kein wesentlicher Verlust an Sprachqualitaet auftritt.

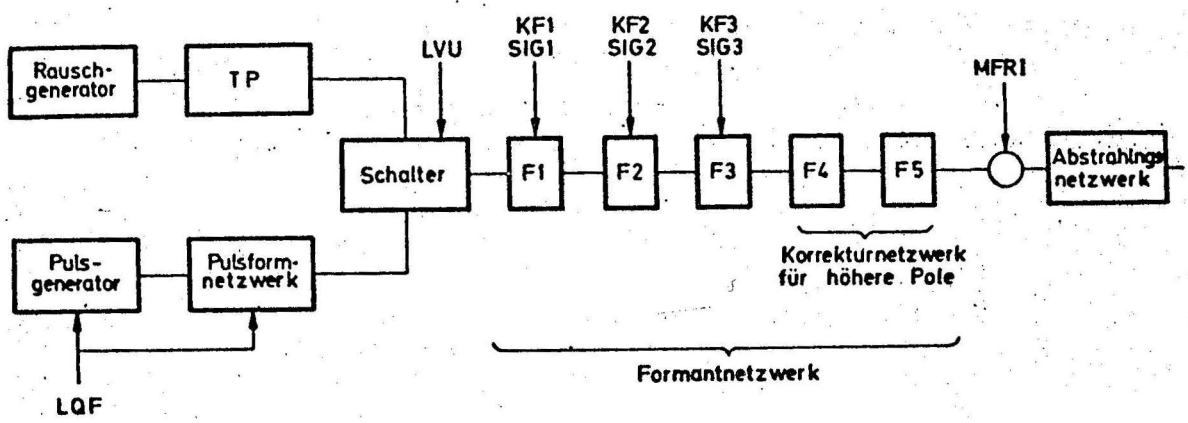


Abb.27, Einfacher Formantvocoder

Eine Amplitudenkontrolle zur Regelung der Gesamtamplitude trennt das Sprachfilter von dem Teil des Formantvocoders,der die Abstrahlung beruecksichtigt. Das Abstrahlungsfilter, das in Kap.5 noch ausfuehrlicher beschrieben wird, benoetigt keine Steuerparameter.

Stimmhaft- Stimmlosigkeit -----	2 bit/20 ms
Pitchfrequenz -----------	3 bit/20 ms
Amplitudenwert -----------	6 bit/20 ms
1. Formantfrequenz ----------	5 bit/20 ms
2. Formantfrequenz ----------	4 bit/20 ms
3. Formantfrequenz ----------	3 bit/20 ms
	23 bit/20 ms

Tabelle 4, Codierung des Formantvocoders

Bei dem Formantvocoder nach Abb.27 muessen ausser der Pitchfrequenz und der Stimmhaft- Stimmlos- Entscheidung, die ja auch beim Kanalvocoder benoetigt werden, nur vier weitere Parameter beruecksichtigt werden: Der Effektivwert der Amplitude und die Frequenzen der ersten drei Formanten. Zur Co-

dierung eines Parametersatzes werden nach den vom Verfasser durchgefuehrten Untersuchungen insgesamt 23 bit benoetigt, wie aus Tab.4 hervorgeht.

Bei einer Abtastung der Parameter in 20 ms- Abstaenden entspricht das einer Bitrate von 1150 bit/sek und damit einem Kompressionsfaktor von 69.
Aufgrund des quasikonstanten Verlaufs der Steuerparameter laesst sich noch eine effektivere Codierung durchfuehren, die in Kap.7 beschrieben wird.

Ein gewisser Nachteil des Formantvocoders besteht darin, dass die Formantbestimmung recht kompliziert ist. Bisher bekannte Hardwarevorrichtungen, die eine Formantbestimmung in Realzeit ermoeglichen, arbeiten nicht mit der erforderlichen Genauigkeit. Es wurden deshalb vom Verfasser die verschiedensten Formantbestimmungsverfahren untersucht.

Alle Formantbestimmungsalgorithmen haben eins gemein. Sie sind bei den notwendigen Genauigkeitsanforderungen mit den heutigen Realisierungsmoeglichkeiten nur auf einem Digitalrechner programmierbar. Die Rechenzeiten betrugen bei der zur Verfuegung stehenden Anlage (CAE 90-40 bzw. IBM 360-67 im CMS- Betrieb) in der Regel mehr als das 200-fache der Realzeit. Die verschiedenen Formant- und Pitchbestimmungsverfahren sind in Kap.6 ausfuehrlich beschrieben.

Fuer den Syntheseteil eines Formantvocoders, der als hardwaremaessige Vorrichtung in Realzeit arbeitet, gibt es in der Literatur viele Loesungsvorschlaege (/26/,/27/,/28/, /29/,/30/). Der Verfasser hat in Kap.5 eigene Untersuchungen auf diesem Gebiet beschrieben und in Kap.8 Vorstellungen fuer einen neuartigen Formantsynthetisator in Hardwareausfuehrung entwickelt. Die Realzeitsynthese soll deshalb an dieser Stelle nicht weiter behandelt werden.

3.5 Vergleich der verschiedenen Methoden

In Tabelle 5 sind die in diesem Kapitel naeher erlaeuterten Vocoderarten noch einmal einander gegenuebergestellt.

Verfahren	Kompressions-faktor	Aufwand zur Parameterer-mittlung	Aufwand zur Realzeitsyn-these
Sprachsynthese aus Gauss-funktionen	>100 (129)	sehr hoch	sehr hoch
Auto-korrelations-vocoder	< 25 (22)	hoch	mittel
Kanal-vocoder	< 33 (33)	gering	mittel
Formant-vocoder	> 50 (69)	hoch	mittel

Tabelle 5, Vergleich der Vocoderarten

Die in Spalte 1 der Tabelle 5 angegebenen Sprachkompressionsfaktoren stellen eine Schaetzung des Verfassers dar. Die in 3.1, 3.2, 3.3 und 3.4 berechneten Kompressionsfaktoren sind jeweils in Klammern dahinter angegeben.

Vergleicht man die gewonnenen Erkenntnisse mit den Anforderungen, die am Anfang des Kap.3 an das Spracherzeugungssystem gestellt wurden, so zeigt sich, dass der Kanalvocoder und der Formantvocoder die besten Voraussetzungen zur Loesung der gestellten Aufgabe mit sich bringen.

Der wesentlich hoehere Sprachkompressionsfaktor des Formantvocoders erscheint dem Verfasser ein groesserer Vorteil als die leichtere Parameterermittlung beim Kanalvocoder zu sein, zumal ja, wie am Anfang erwaehnt wurde, ein endlicher Wortvorrat durch eine einmalige Analyse geschaffen werden sollte. Aus diesem Grund wird vom Verfasser der Formantvocoder als der Vocodertyp betrachtet, der den gestellten Anforderungen von allen bekannten Vocoderarten am besten gerecht wird.

In den Kapiteln 5, 6, 7 und 8 wird der Formantvocoder ausfuehrlich beschrieben.

4. Hilfsmittel zur Sprachverarbeitung

4.1 Verwendete Rechenanlage

Bei der im wesentlichen verwendeten Rechenanlage handelt es sich um das hybride Rechnersystem HRS 900.

Es besteht aus dem Digitalrechner CAE 90-40 (SDS 930), der ueber das Koppelwerk HKW 900 mit dem Analogrechner RA 770 der Firma Telefunken zusammengeschaltet ist. Der Aufbau der gesamten Anlage ist in Abb.28 dargestellt.

Der Rechner verfuegt ueber einen Kernspeicher von 16 K-24- bit- Worten.

An die Zentraleinheit sind zwei Ein- Ausgabekanaele, der W- Kanal und der Y- Kanal angeschlossen, die im Zeitmultiplexbetrieb arbeiten.

An den Y- Kanal sind ein Schnelldrucker und eine Platteneinheit angeschaltet. Die Platteneinheit hat eine Kapazitaet von 24 Mio. bit und eine maximale Lesegeschwindigkeit von 960 kbit/sek.

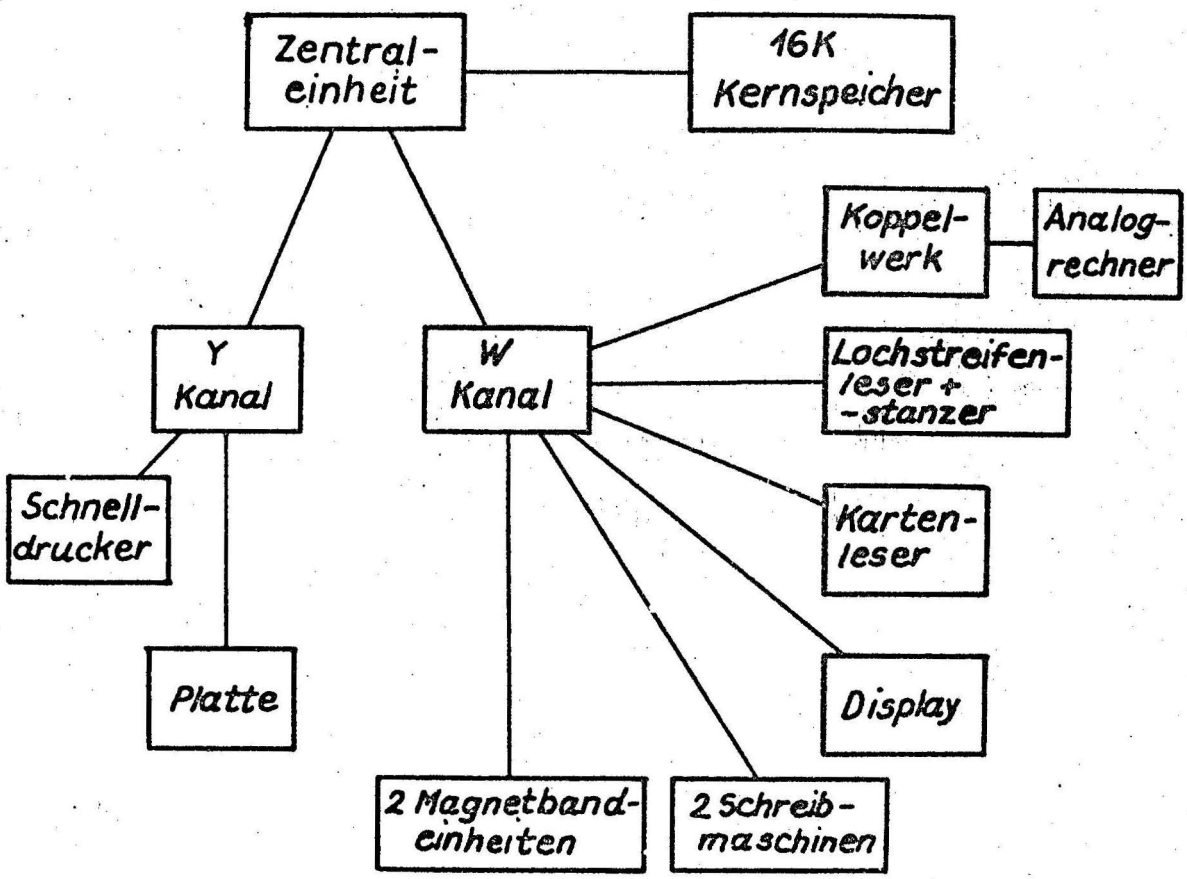


Abb.28, Aufbau der verwendeten Rechenanlage

An den W- Kanal sind zwei Magnetbandeinheiten, ein Kartenleser, zwei Fernschreibmaschinen, ein Lochstreifenstanzer und -leser, die Displayeinheit SDS 9185 und das Koppelwerk HKW 900 angeschlossen.
Der Analogrechner RA 770 ist an das Koppelwerk HKW 900 geschaltet.

Ausser dieser Realzeitanlage, die im folgenden noch genauer beschrieben wird, stand eine IBM 360-67 zur Verfuegung, die vorwiegend im CMS- Betrieb benutzt wurde.

Digitalrechner

Der Digitalrechner CAE 90-40 weist eine Reihe von Merkmalen auf, die ihn zur Bearbeitung von Realzeitaufgaben besonders geeignet machen. Dazu gehoeren:

1. Kurze Zyklus- und Speicherzugriffszeiten (1.75 bzw. 0.8 µs)
2. Automatischer blockweiser Datentransfer zwischen Kernspeicher und Peripherie (Interlace)
3. Ein- und Ausgabe von Steuersignalen getrennt von den normalen Ein/ Ausgabe- Datenkanaelen (PIN/POT- und EOM- Befehle)
4. Abfrageleitungen (sense lines) zur programmierbaren Abfrage angeschlossener Peripherieeinrichtungen auf das Vorhandensein von Einzelbit- Signalen
5. Unterbrechungsleitungen (Interrupt- lines) mit verschiedenen und vom Programm her aktivierbaren Vorrangstufen zur Programmunterbrechung durch Einzelbit- Signale von Peripherieeinrichtungen.

Zur Programmierung des Digitalrechners stehen im wesentlichen zwei Assembler (SYMBOL und META-SYMBOL), ein FORTRAN II und ein REAL-TIME-FORTRAN- Compiler zur Verfuegung.
Das FORTRAN II hat viele Eigenschaften des FORTRAN IV. Es fehlen jedoch doppelt- genaue, komplexe und logische Operationen.
In REAL-TIME-FORTRAN koennen Interrupts angeschlossen und in dem Zusammenhang auch rekursive Unterprogramme geschrieben werden. FORTRAN- Statements und Maschinenbefehle im Assembler- Code duerfen beliebig miteinander gemischt werden. Die Programmierung sowohl der Displayeinheit als auch des Hybridsystems kann vollstaendig mit FORTRAN- Anweisungen durchgefuehrt werden.

Koppelwerk

Das Koppelwerk enthaelt fuer die Verarbeitung der analogen Spannungsverlaeufe einen 16- Kanal- Multiplexer mit Abtast-Halteglied. Diesem ist der Analog- Digital- Umsetzer (ADU) VT 13-AB der Firma ADAGE nachgeschaltet. Er gestattet 14 bit in 5 µs umzusetzen. Bei der Verwendung der in REAL-TIME-FOR-

TRAN aufrufbaren Hilfsprogramme laesst sich in der Praxis jedoch nur eine Abtastzeit von 4 ms erreichen.

Zur Digital- Analog- Umsetzung (DAU) stehen 10 Umsetzer vom Typ DAS 900 der Firma TELEFUNKEN zur Verfuegung. Sie setzen 14 bit in maximal 10 µs um. Die DAUs haben die folgenden vier verschiedenen Betriebsarten:

Umsetzen
Multiplizieren
Extrapolieren, fein
Extrapolieren, grob

In der Betriebsart 'Multiplizieren' kann vom Analogrechner eine beliebige Referenzspannung aufgeschaltet werden, so dass die DAUs als elektronische Potentiometer arbeiten, deren Einstellung vom Digitalrechner gesteuert werden kann.

Analogrechner

Der Analogrechner RA 770 ist ein sog. hybrider Analogrechner. Er verfuegt ueber einen Digitalzusatz. Auf 24 Magazinplaetzen sind in einer beliebigen Verteilung Steckkarten mit einer jeweils groesseren Anzahl von digitalen Elementen, wie Flipflops, Monoflops, Schieberegister, Zaehler, Inverter, NAND- und NOR- Gliedern einsetzbar. Die Ein- und Ausgaenge dieser Elemente enden an den Buchsen eines Digitalprogrammierfeldes. Dort werden die Digitalelemente mit Schnueren und Kurzschlusssteckern, aehnlich wie die Elemente des Analogprogrammierfeldes, programmiert.

Vom Digitalprogrammierfeld lassen sich beispielsweise die Integrierer und Komparatorschalter auf dem Analogsteckbrett, aber auch die Betriebsarten, wie PAUSE, RECHNEN und HALT steuern.

Die Ausgaenge der Komparatorverstaerker des Analogprogrammierfeldes enden ebenfalls auf dem Digitalprogrammierfeld.

Das Analogprogrammierfeld ist in 10 Felder unterteilt. Jedes Feld enthaelt im wesentlichen:

3 Summierer/Integrierer/Speicher
2 Summierer
2 Handpotentiometer
5 Servopotentiometer
1 Nichtlineares Netzwerk, wie
Parabelmultiplizierer oder
Funktionsgeber

Der Analogrechner arbeitet mit einer Referenzspannung von 10 V und einer Rechengenauigkeit von 0.0001.

4.2 A/D- und D/A- Umsetzung (/9/)

Tiefpassfilterung

Die Verarbeitung von Sprache auf dem Digitalrechner kann nur dann durchgefuehrt werden, wenn die Information, hier die Sprache, in digitalisierter Form vorliegt. Das geschieht dadurch, dass man zunaechst eine Spannung erzeugt, deren Verlauf proportional dem Druck an einem Ort des Schallfeldes ist, sie geeignet filtert und schliesslich diese Spannung zu aequidistanten Zeitpunkten abtastet. Das Resultat ist eine Zahlenfolge, deren Glieder zeitlich fest zueinander in Beziehung stehen und deren Betraege der Amplitude des Drucks in dem Schallfeld entsprechen.

Durch die Abtastung des Analogsignals erfolgt eine Beschneidung des Frequenzbandes des Originalsignales. Nach dem Shannon'schen Abtasttheorem muss die Abtastfrequenz mindestens doppelt so gross wie die hoechste zu verarbeitende Frequenz sein. Es wurde eine Abtastfrequenz von 10 kHz gewaehlt. Dieser Wert wird auch in der Literatur am haeufigsten genannt.

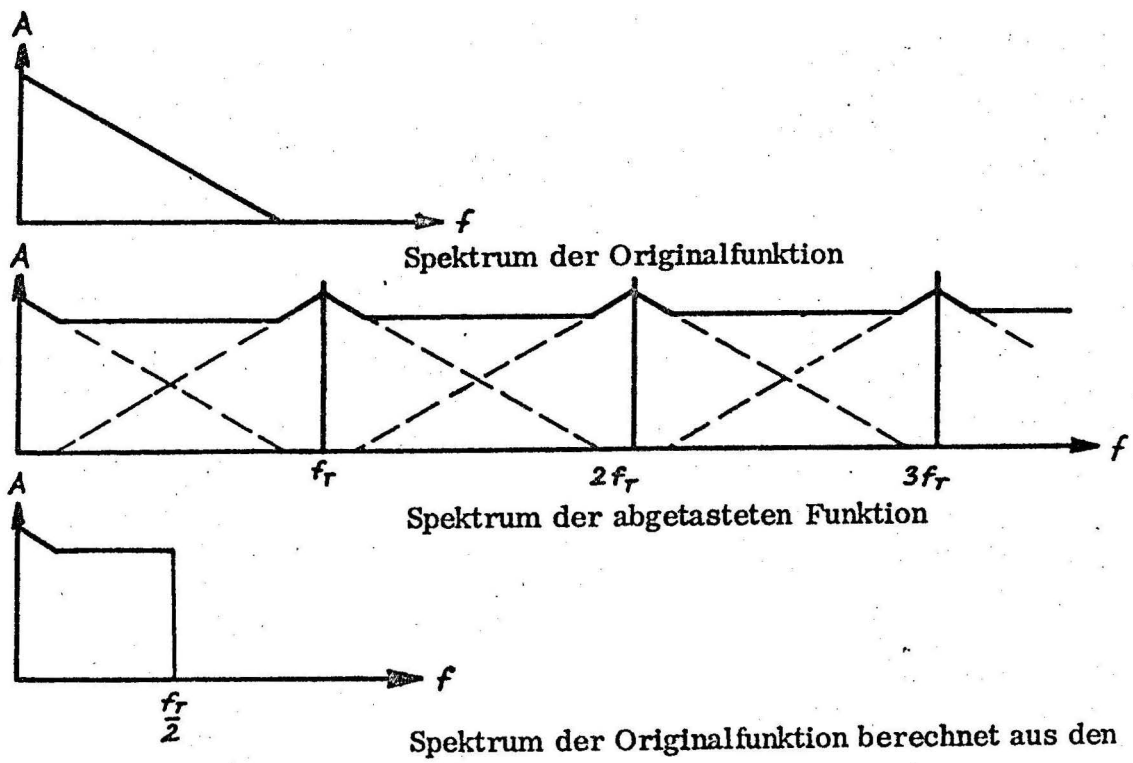


Abb.29, Spektren des ungefilterten Signals

Durch eine hoehere Abtastfrequenz wird nur bedingt eine hoehere Sprachqualitaet erreicht, da die Schwierigkeiten, die bei einer Simulation aufgrund der endlichen Registerlaenge eines Digitalrechners auftreten, auch groesser werden.

Das Spektrum eines abgetasteten Vorgangs weist Ueberlagerungen durch die Faltung an den ganzen Vielfachen der Abtastfrequenz auf. Dadurch kann das urspruengliche Spektrum nicht mehr aus der abgetasteten Zeitfunktion gewonnen werden.

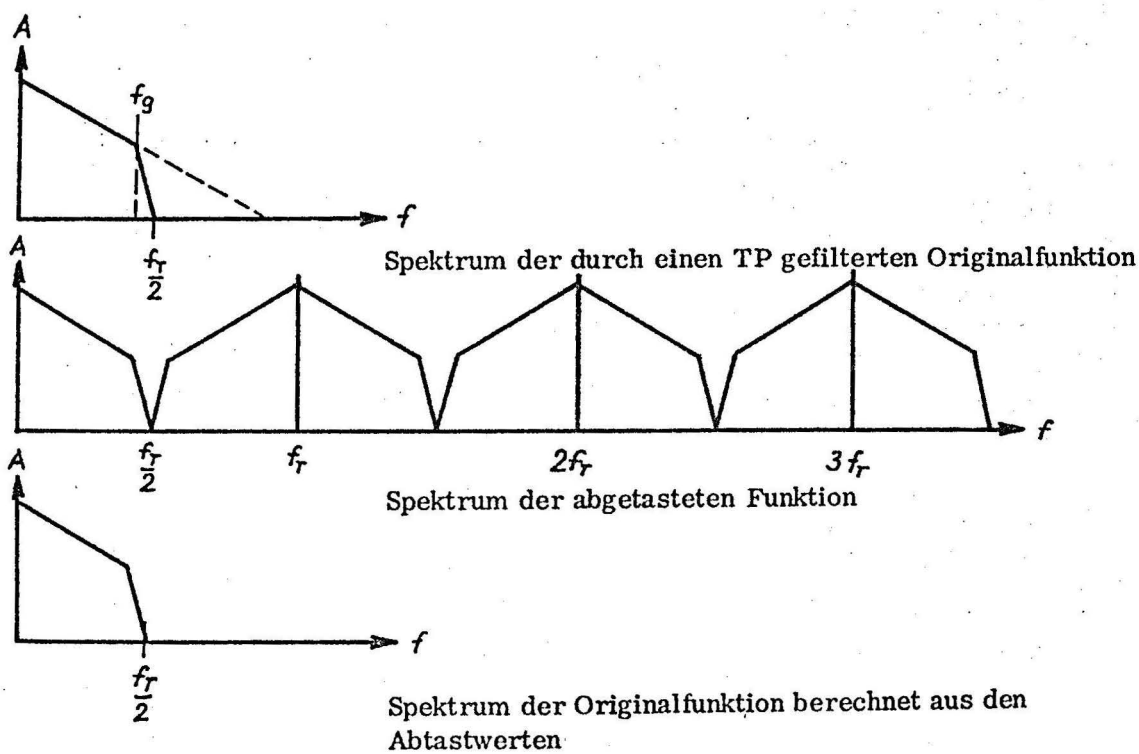


Abb.30, Spektren des gefilterten ... nals

Abb.29 zeigt ein Beispiel fuer das Spektrum einer analogen Zeitfunktion, das der mit $f_T = 1/T$ abgetasteten Zeitfunktion entspricht, und schliesslich das Spektrum der Originalfunktion, wie es aus der abgetasteten Funktion berechnet werden wuerde. Man sieht, dass das berechnete Spektrum nicht mehr mit dem Spektrum der Originalfunktion uebereinstimmt.

Braucht man fuer die Weiterverarbeitung des Signals den Spektralverlauf des urspruenglichen Spektrums, muss man durch Tiefpassfilterung des Signals dafuer sorgen, dass die Ueberlagerungen vernachlaessigbar klein werden. Abb.30 zeigt die entsprechenden Spektren zu Abb.29 fuer den Fall, dass die Originalfunktion mit einem Tiefpass der Grenzfrequenz f_g

gefiltert wurde. Man sieht, dass sich jetzt aus der abgetasteten Zeitfunktion das Spektrum der Originalfunktion weitaus genauer berechnen laesst, als im anderen Falle. Die Mikrophonspannung wurde deshalb durch einen Tiefpass gefiltert. Bei dem Tiefpass handelt es sich um ein dreipoliges Tschebyscheff- Filter mit einer Grenzfrequenz f_g =3 kHz und einer Welligkeit von 1%.

Die Aufnahmen wurden in einem schalltoten Raum durchgefuehrt, um akustische Reflektionen auszuschalten.

Frequenztransformation

Die Abtastung des gefilterten Signals erfolgt mit dem Hybridsystem HRS 900. Wie schon oben beschrieben wurde, gestattet das zur Verfuegung stehende Hybridsystem mit seiner in FORTRAN geschriebenen Software nur Abtastfrequenzen bis zu 250 Hz. Deshalb muss die analoge Zeitfunktion mindestens um den Faktor 40 in der Frequenz transformiert werden, wenn eine Tastfrequenz von 10 kHz erreicht werden soll. Die Frequenztransformation wird mit einem Analog- Magnetbandgeraet durchgefuehrt. Die analoge Zeitfunktion wird zunaechst mit einer hohen Bandgeschwindigkeit aufgenommen und dann mit einer geringen Geschwindigkeit wiedergegeben und abgetastet. Dadurch multipliziert sich die Abtastfrequnez des Hybridsystems mit der Untersetzung des Bandgeraetes.

Zur Untersetzung wurde das Magnetbandgeraet 7001 von Brueel & Kjaer verwendet. Es handelt sich dabei um einen Zweikanal- Analogspeicher in Frequenzmodulationstechnik fuer Signale von 0Hz (Gleichspannung) bis 20 kHz. Vier Bandgeschwindigkeiten ermoeglichen Frequenz- bzw. Zeittransformationen in acht Stufen von 1:40 bis 40:1. Das ermoeglicht zusammen mit der maximalen Abtastrate des Hybridsystems Abtastfrequenzen bis zu 10 kHz.

Abtastung

Die Analyse der Sprache kann wegen der Kompliziertheit der zugehoerigen Programme nicht in Realzeit und zunaechst auch nicht in der vierzigfachen Realzeit erfolgen. Deswegen wuerden die Abtastdaten schneller anfallen, als sie verarbeitet werden koennen. Man muss sich deshalb zunaechst auf das Abtasten beschraenken.

Da der verfuegbare Speicherplatz des Digitalrechners nur 16 K umfasst, aber pro Sekunde abzutastender Sprache bereits 10000 Werte anfallen, muessen die Daten auf ein anderes externes Speichermedium transportiert werden, und zwar zweckmaessigerweise auf Magnetband oder Platte.

Um die Probleme, die bei der Abtastung auftreten, besser verstehen zu koennen, muss hier kurz etwas zum synchronisierten Datentransfer des Hybridsystems gesagt werden.

Der Hybridrechner arbeitet in zwei Modi, im Modus CONTROL und im Modus WORK. Der Modus CONTROL entspricht der Analogrechnerstellung PAUSE und der Modus WORK der Analogrechnerstellung RECHNEN. Waehrend des Modus WORK finden in

aufeinanderfolgenden Zyklen hintereinander die folgenden Funktionsablaeufe statt:

1. Datentransfer DR → AR
2. Datentransfer AR → DR
3. Rechnen (im Digitalrechner)
4. Warten bis Zyklusende

Die Zeit eines solchen Zyklus wird auch Frametime genannt und kann per Programm vom Digitalrechner eingestellt werden. Da in jedem Zyklus der Datenaustausch zwischen Analog- und Digitalrechner nur einmal stattfindet, ergibt sich, dass die Frametime gleich dem Kehrwert der Abtastfrequenz sein muss. Bei einer gewuenschten Abtastfrequenz von 10 kHz betraegt die Frametime 4 ms, wenn man eine vierzigfache Frequenztransformation durch das Analogmagnetbandgeraet beruecksichtigt. In diesen 4 ms wird also ein Analogwert abgetastet und in den Kernspeicher des Digitalrechners gebracht. Die Zeit reicht aber nicht mehr dazu aus, den Abtastwert auf Magnetband oder Platte zu schreiben. Der Abtastwert muesste in dem Fall als ein Block geschrieben werden. Der zu einem Block gehoerende Anlauf- und Stoppschritt betraegt beim Magnetband zusammen aber bereits schon 25 - 30 ms, bzw. die Zugriffszeit der Platte ca 20 ms. Die Abtastung muss deshalb so organisiert werden, dass zunaechst ein Zahlenfeld von etwa 8000 Zellen im Kernspeicher aufgefuellt und die Abtastung danach durch den Ruecksprung vom Modus WORK in den Modus CONTROL unterbrochen wird. Im Modus CONTROL wird das Zahlenfeld auf Magnetband geschrieben und anschliessend der naechste Analogblock im Modus WORK abgetastet. Ist die Abtastung eines Blocks beendet, muss auch das Analogmagnetbandgeraet angehalten und erneut gestartet werden, wenn der naechste Block abgetastet werden soll.

Markierung /31/

Es besteht die Gefahr, dass zwischen Starten und Stoppen des Analogmagnetbandgeraetes ein Teil der Information verlorengeht. Deshalb werden Anfang und Ende eines jeden Analogblocks auf einer zweiten Spur des Magnetbandes durch einen Puls markiert. Die Markierung kann durch Betaetigung einer Taste von Hand erfolgen oder vom Rechner durch ein Markierprogramm. Es ist dabei lediglich zu beachten, dass die Laenge eines Analogblockes nicht den Wert ueberschreitet, der 8000 Abtastwerten entspricht, da nicht mehr als 8000 Zellen im Kernspeicher fuer die Abtastwerte reserviert worden sind.

Wird das Magnetbandgeraet gestartet und befindet sich der Hybridrechner im Modus WORK, so leitet der erste Markierpuls der jetzt vom Band gelesen wird, ueber eine Interruptleitung den Abtastvorgang ein, waehrend der zweite Markierpuls, der gelesen wird, ueber eine zweite Interruptleitung die Abtastung beendet. Das Bandgeraet wird selbsttaetig angehalten und muss um einige Zentimeter zurueckgespult werden, damit es beim erneuten Start die letzte Endmarke jetzt als Anfangsmarke lesen kann.

Digital- Analog- Umsetzung

Bei der Digital- Analog- Umsetzung, die aus den Abtastwerten synthetischer Sprache wieder die kontinuierliche Zeitfunktion auf dem Analog- Magnetband erzeugen soll, tritt ein aehnliches Problem, wie bei der Abtastung auf. Die Frametime, die jetzt wiederum 4 ms beraegt, reicht nicht aus, um den Wert vom Digital- Magnetband oder der Platte zu lesen. Deshalb wird entsprechend zur Abtastung im Modus CONTROL ein Block im Kernspeicher mit den synthetisch erzeugten Abtastwerten gefuellt. Im Modus WORK wird durch die Anfangsmarke auf dem Analog- Magnetband ein Interrupt erzeugt, der die Ausgabe der Analogwerte einleitet. Am Ende eines ausgegebenen Blocks wird auf die zweite Spur zusaetzlich eine Marke gesetzt, die das Ende des Analogblocks auf Spur 1 anzeigt. Im Modus CONTROL wird dann der naechste Block vom Digitalmagnetband gelesen und das Analog- Magnetband um einige Zentimeter zurueckgespult, so dass bei einer erneuten Ausgabe die vom Rechner unmittelbar vorher geschriebene Endmarke wieder als Anfangsmarke gelesen werden kann.

Es wird im Augenblick daran gearbeitet, unter Ausnutzung der Interlacetechnik und der hohen seriellen Lese- und Schreibgeschwindigkeit der Platte eine A/D- und D/A- Umsetzung aufzubauen, die 10 kHz Samplingfrequenz in Realzeit zu verarbeiten vermag.

4.3 Visible-Speech-Darstellung (/32/,/33/)

Die Qualitaet von Sprache laesst sich i.A. nur nach Demonstration akustischer Hoerproben beurteilen. In vielen Faellen, wie Veroeffentlichungen, in denen es nicht immer moeglich ist, dem Druck z.B. eine Schallplatte beizulegen, benutzt man die Visible- Speech- Darstellung, um einen Eindruck der synthetisierten Sprache zu vermitteln. Die optische Darstellung der Sprache ist insbesondere dort sehr gut zu gebrauchen, wo es sich um Vergleiche, z.B. der Originalsprache mit der synthetisierten Sprache handelt. Ausserdem lassen sich aus der Visible- Speech- Darstellung Rueckschluesse auf die Struktur der Sprache ziehen.

Die Visible- Speech- Darstellung ist eine raeumliche Darstellung des Verlaufes des Kurzzeitspektrums ueber der Zeit. Die Abb.31 skizziert die Zuordnung der Koordinaten der

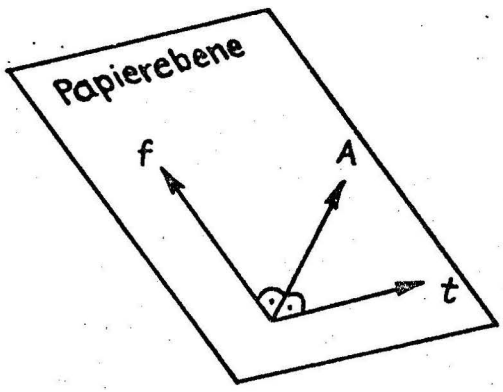


Abb.31, raeumliche Achsen bei der
Visible- Speech- Darstellung

raeumlichen Darstellung zur Papierebene. Darin bedeuten:

f die Frequenzachse
A die Amplitudenachse
t die Zeitachse

Die Amplitudenachse, die senkrecht zur Bildebene steht, wird auf dem Papier durch verschiedene Schwaerzungsgrade dargestellt.
Die im folgenden beschriebene Visible- Speech- Darstellung wird durch ein Programm auf dem Digitalrechner CAE 90 - 40 dargestellt. Als Ausgabemedium dient der Schnelldrucker.
Die verschiedenen Schwaerzungsgrade werden durch Vierfachdruck auf dem Schnelldrucker erzeugt. Es wird dabei ein Graukeil verwendet, der 9 Abstufungen aufweist. GERULL /32/ hat verschiedene Zeichenkombinationen zur Erzeugung eines guten Graukeils untersucht. Die Tabelle 6 gibt die verwendeten Zeichen fuer die einzelnen Graustufen und Druckschritte an, die den besten Graukeil auf dem zur Verfuegung stehenden

Schnelldrucker erzeugten. Die weissen Felder in Tabelle 6 stellen Blanks dar.

Helligkeitsgrad	Nummer des Druckschrittes 1.	2.	3.	4.
1				
2	'			
3	/			
4	Y			
5	O			
6	U	I		
7	O	X		
8	O	$		
9	=	O	$	B

Tabelle 6, Drucktypen fuer Vierfachdruck

Das Kurzzeitspektrum einer Zeitfunktion ergibt sich dadurch, dass man das Spektrum aus der Zeitfunktion s(t), die durch das Zeitfenster h(t) betrachtet wird, berechnet.
Das Zeitfenster h(t) hat die Laenge T und es gilt:

h(t)=0 fuer t<-T/2 und t>T/2

Dann berechnet sich das Kurzzeitspektrum zu:

$$S(\omega,t) = \left[\int_{-\frac{T}{2}}^{\frac{T}{2}} f(\tau) \cdot h(\tau - t)\, e^{-s\tau}\, d\tau \right]_{s=j\omega} \quad (39)$$

bzw. mit $F(s) = \mathcal{L}[f(t)]$ und $H(s) = \mathcal{L}[h(t)]$ ergibt sich

$$S(\omega,t) = \left[F(s) * H(s) \right]_{s=j\omega} \quad (40)$$

F(s) stellt das Sprachspektrum dar. Es reicht, wie aus der Abb.30 hervorgeht, bis zur Eckfrequenz des verwendeten Tiefpass, also bis 3 kHz.
Das Glied H(s) stellt das Spektrum des Zeitfensters dar.
Unter der Annahme eines rechteckfoermigen Verlaufs des Zeitfensters, d.h. :

fuer -T/2 < t < T/2 ist h(t)=1 und sonst ist h(t)=0

ergibt sich der Spektralverlauf: $$|H(s)|_{s=j\omega} = T \frac{\sin \frac{\omega T}{2}}{\frac{\omega T}{2}} \quad (41)$$

Das Spektrum des Zeitfensters hat in diesem Fall den Verlauf einer sin(x)/x - Funktion. Die effektive Fensterbreite im Frequenzbereich sei durch die erste Nullstelle der sin(x)/x-Funktion in positiver und negativer Richtung begrenzt. Dann ist

die Fensterbreite im Zeitbereich: $D_t = T$ (42)
die Fensterbreite im Frequenzbereich $D_f = 2\omega_g = \frac{4\pi}{T}$ (43)

Aus Gl.(42) und Gl.(43) ergibt sich, dass mit der Laenge des Zeitfensters stets ein Kompromiss zwischen Frequenzaufloesung und Zeitaufloesung geschlossen wird: Ein schmales Fenster im Zeitbereich bedeutet eine hohe Zeitaufloesung, aber nur eine geringe Frequenzaufloesung. Dagegen bedeutet ein schmales Fenster im Frequenzbereich eine hohe Frequenzaufloesung, aber nur eine geringe Zeitaufloesung.

Wie aus Gl.(18) und Gl.(19) hervorgeht, enthaelt die Sprache Bestandteile, die von der Quelle, vom Vokaltrakt und von der Abstrahlung herruehren. Im Kurzzeitspektrum findet sich daher eine Grobstruktur, die die Uebertragungseigenschaften des Vokaltraktes und den Einfluss der Abstrahlung wiederspiegelt, ueberlagert von einer Feinstruktur, die von der Quelle herruehrt. Durch die Wahl eines geeigneten Zeitfensters kann Einfluss darauf genommen werden, ob im Kurzzeitspektrum die Feinstruktur erscheinen soll, oder nicht. Die Abb.32a bis 32e demonstriert die Visible- Speech- Darstellung des Wortes HAWAII mit Zeitfenstern verschiedener Laenge T.

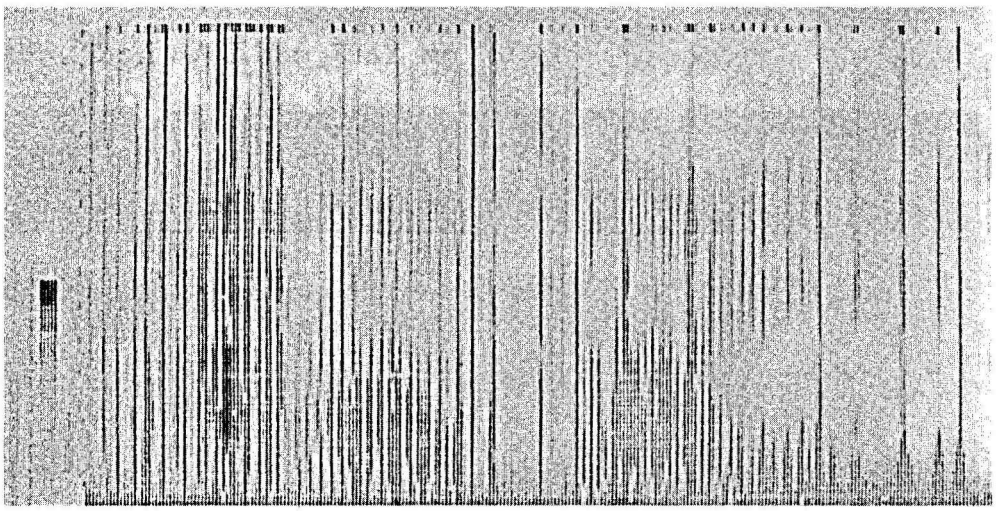


Abb.32a, HAWAII: T=3 ms

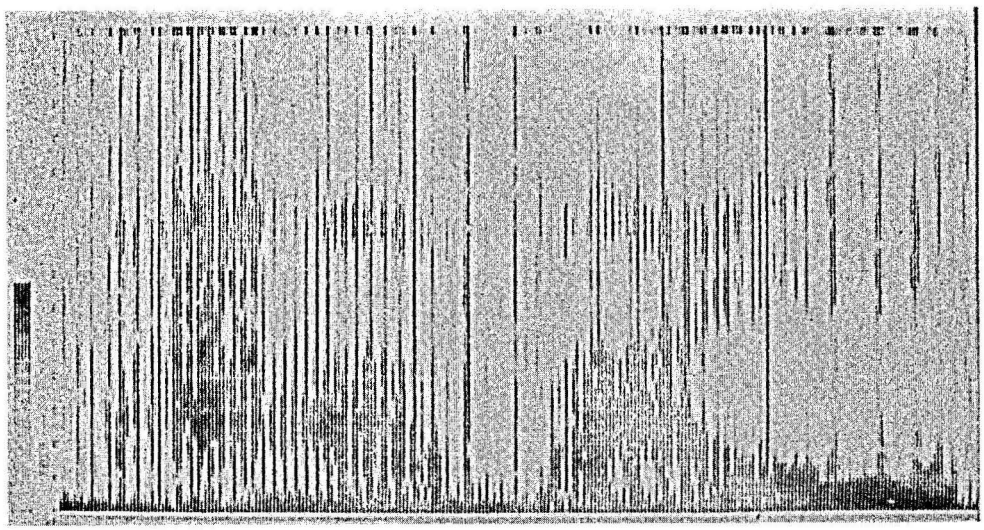


Abb.32b, HAWAII: T=6 ms

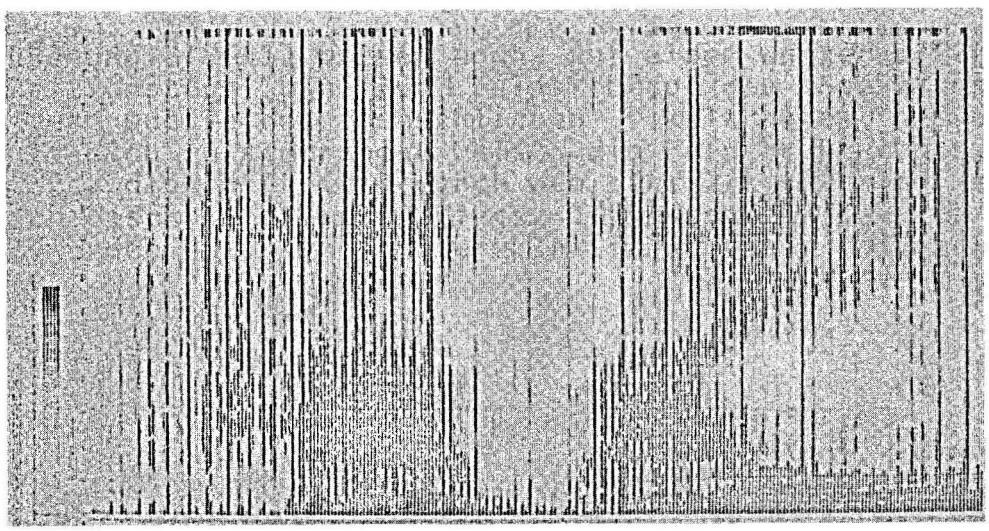


Abb.32c, HAWAII: T=12 ms

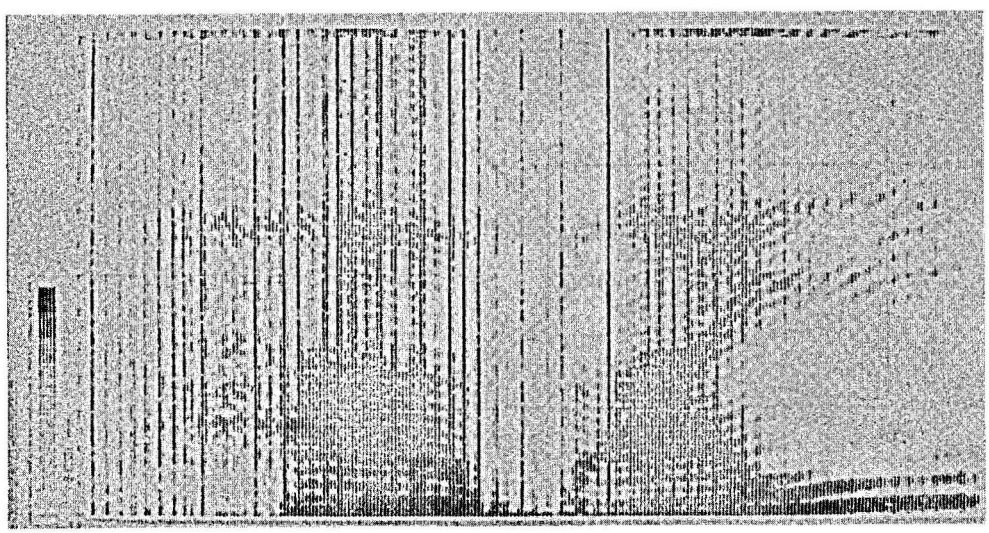


Abb.32d, HAWAII: T=24 ms

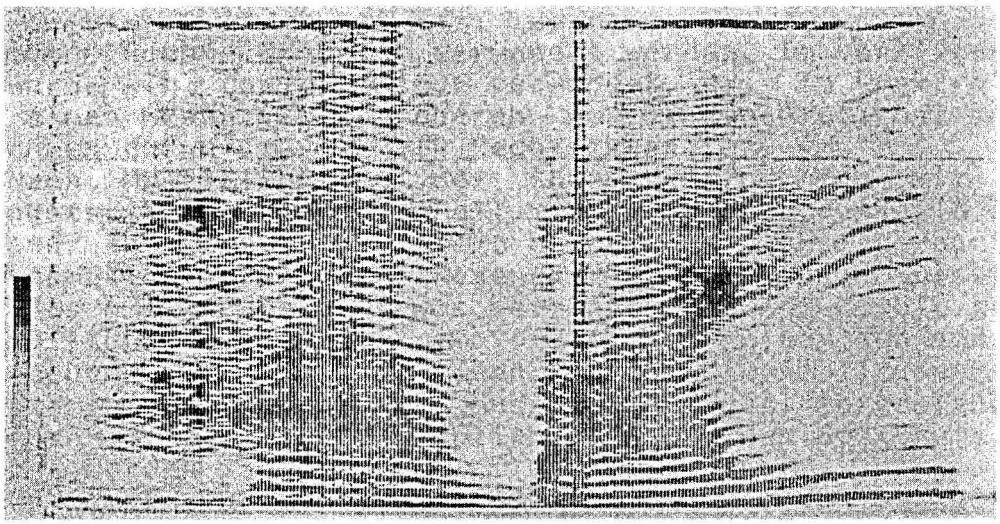


Abb.32e, HAWAII: T=51.2 ms

Gl.(40) und Gl.(41) weisen noch auf einen weiteren Einfluss des Zeitfensters hin. Zur Feinstruktur des Spektrums, das von der Quelle des menschlichen Vokaltraktes herruehrt, addiert sich in stoerender Weise die Feinstruktur des Zeitfensters, die in Gl.(41) durch den sin(x)/x - Verlauf mathematisch beschrieben wird. Dieser Einfluss kann dadurch reduziert werden, dass geeignetere Zeitfenster als das oben be-

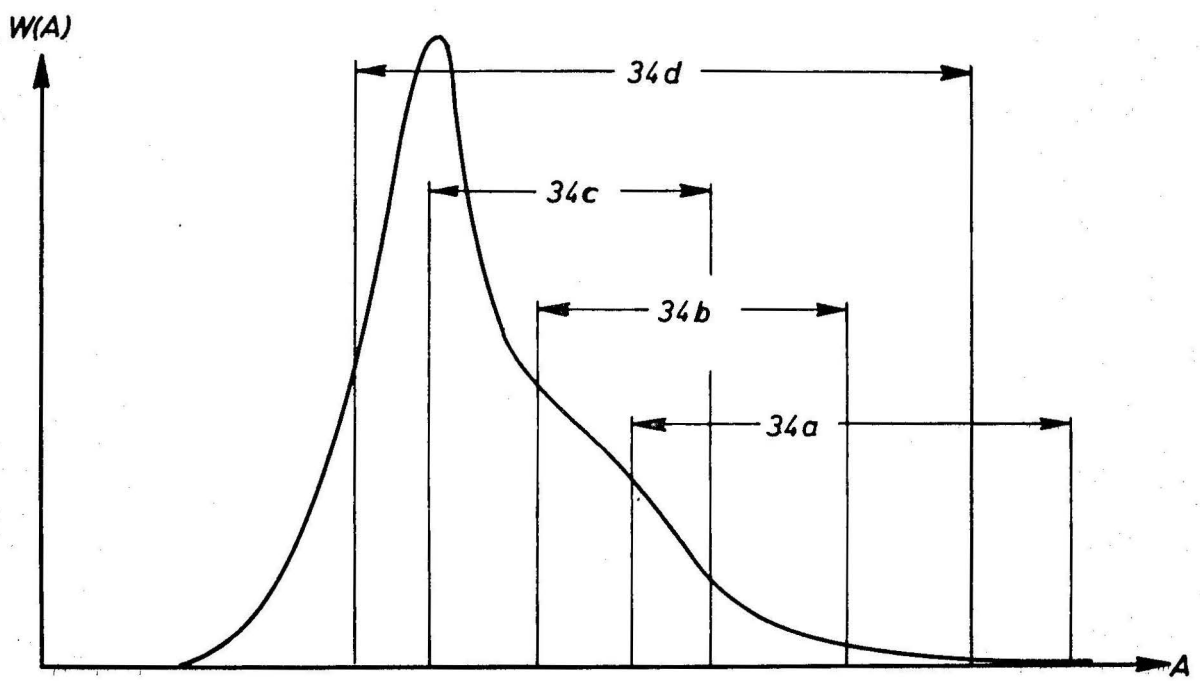


Abb.33, Verteilungskurve der Amplitudenwerte des Spektrums fuer BODEN

schriebene Rechteckfenster verwendet werden. Im vorliegenden Fall wurde ein $\cos^2$ -Fenster verwendet, das im Spektralbereich einen wesentlich rascheren Amplitudenabfall bei hohen Frequenzen aufweist, als ein Rechteckfenster.

Nach Multiplikation der Sprachzeitfunktion mit dem Zeitfenster koennen in dem Fall, in dem die Sprache in Form diskreter Abtastwerte vorliegt durch die diskrete Fouriertransformation und anschliessende Betragsbildung diskrete Werte des Amplitudenspektrums berechnet werden. Um die grossen Amplitudenunterschiede im Spektrum bei niedrigen und hohen Frequenzwerten auszugleichen, wurden die Amplitudenwerte logarithmiert. Anschliessend muss der Graukeil in geeigneter Weise den logarithmierten Amplitudenwerten des Spektrums zugeordnet werden /33/.

Es werden zunaechst saemtliche Spektren des darzustellenden Wortes berechnet und eine Verteilungskurve fuer die Amplituden der Spektren aufgestellt. Den Verlauf der Verteilungskurve fuer das Wort BODEN zeigt die Abb.33.
Fuer vier Beispiele, die in Abb.34a bis 34d wiedergegeben sind, wurden in Abb.33 die Graukeilgrenzen eingezeichnet.

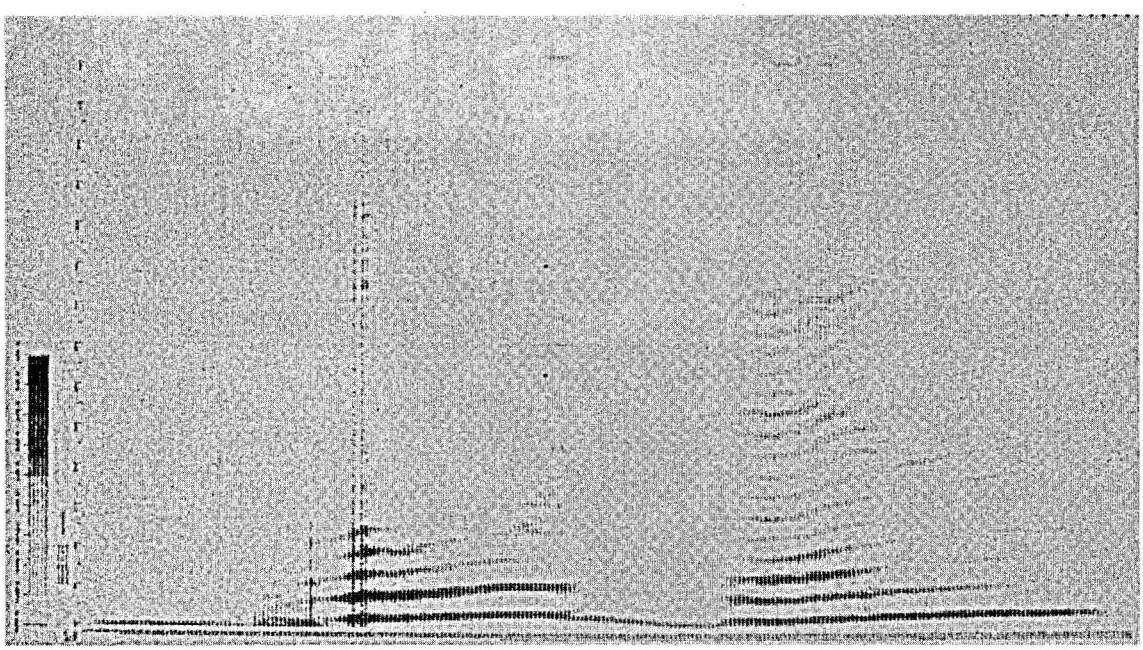


Abb.34a, Lage des Graukeils siehe Abb.33

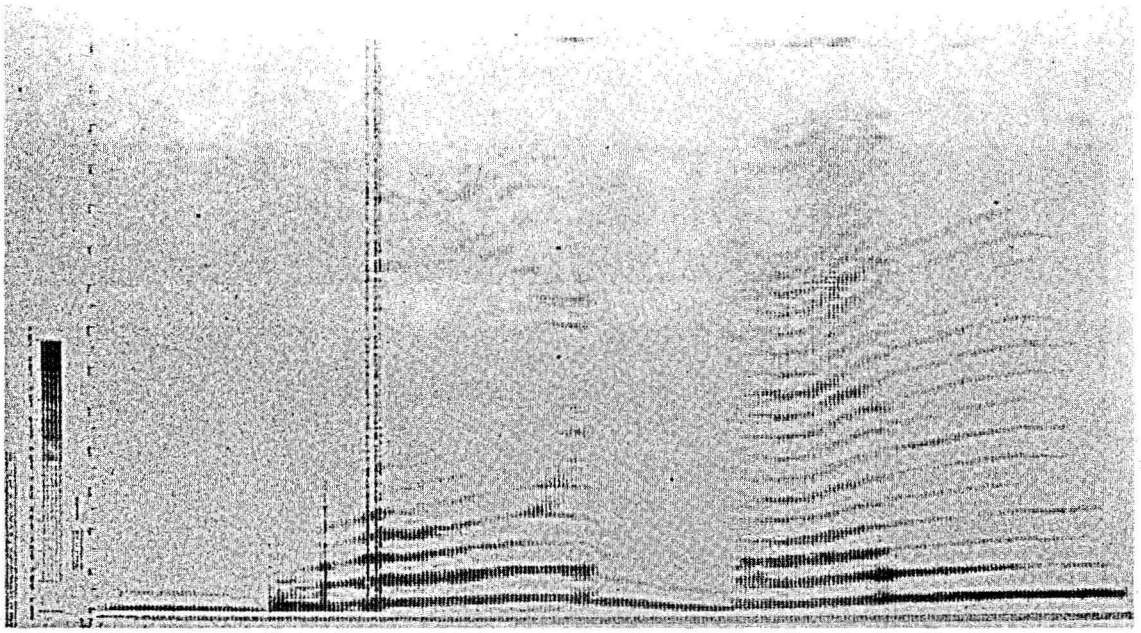


Abb.34b, Lage des Graukeils siehe Abb.33

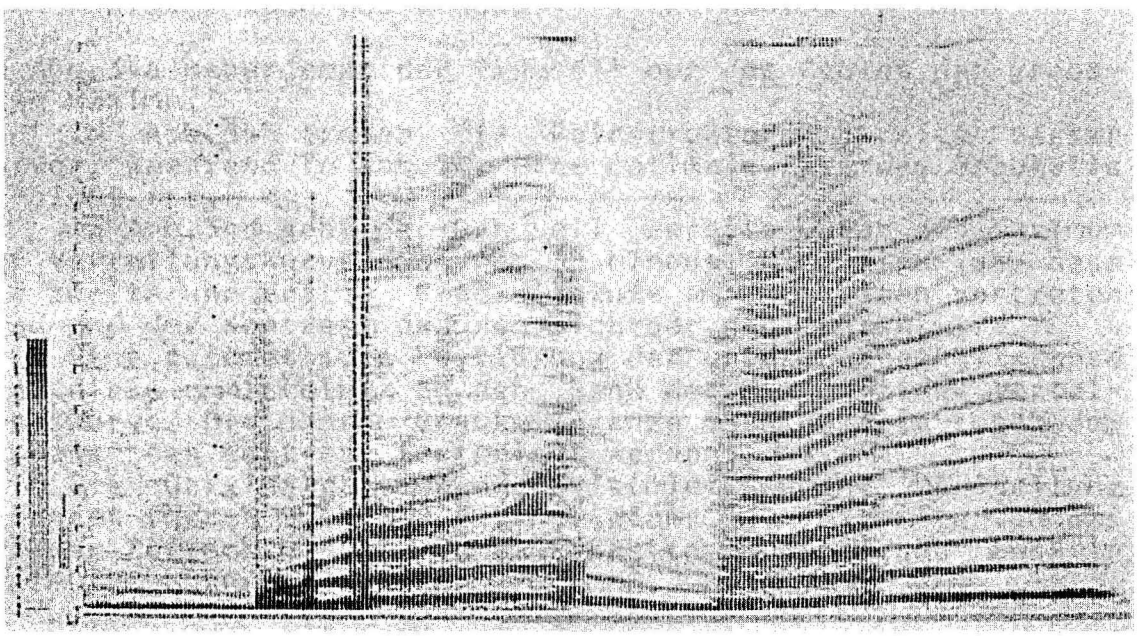


Abb.34c, Lage des Graukeils siehe Abb.33

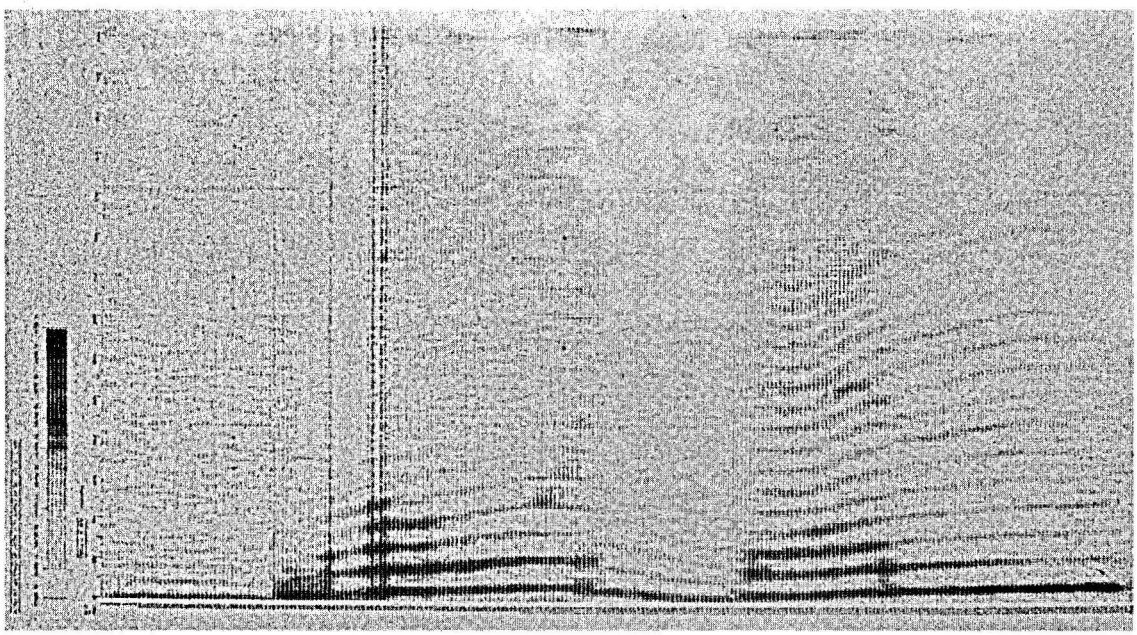


Abb.34d, Lage des Graukeils siehe Abb.33

In Abb.34a ueberdeckt der Graukeil nur das Gebiet der groessten Maxima.

In Abb.34b treten die Feinstrukturen bereits besser hervor, waehrend in Abb.34c eine optimale Lage des Graukeils erreicht ist.

In Abb.34d geht der Graukeil bereits ueber das Maximum der Verteilungskurve nach Abb.33 hinaus. Die Folge ist, dass die zweite und dritte Graukeilstufe am staerksten vertreten sind und der Kontrast dadurch sichtbar schlechter wird.

Eine automatische Festlegung der unteren Graukeilgrenze legt diese unmittelbar an den Rand des Maximums der Verteilungskurve. Die obere Graukeilgrenze muss dann je nach dem gewuenschten Kontrast festgelegt werden.

Die Originalgroesse der Visible- Speech- Darstellung betraegt fuer 1 sek Sprache bei einer Schrittweite von 5ms auf der Zeitachse 85 * 60 cm (Breite * Hoehe) und muss in zwei Druckbahnen hergestellt werden.

5. Syntheseteil des Formantvocoders

5.1 Serien- und Parallelschaltung

Der Aufbau eines Formantvocoders wurde bereits unter 3.4 kurz behandelt. Es wurde mehrfach erwaehnt, dass die Uebertragungseigenschaften des Vokaltraktes allgemein durch Gl.(5) und speziell fuer die Erzeugung von Vokalen durch Gl.(12) beschrieben wird. Gl.(12) lautet:

$$H(s) = \prod_{n=1}^{\infty} \frac{s_n \cdot s_n^*}{(s-s_n)\cdot(s-s_n^*)} \qquad (12)$$

Durch Partialbruchzerlegung laesst sich die Gleichung folgendermassen umformen:

$$H(s) = \sum_{n=1}^{\infty} C_n (s-s_{2n}) \frac{s_n \cdot s_n^*}{(s-s_n)\cdot(s-s_n^*)} \qquad (44)$$

Aus Gl(44) geht hervor, dass sich die Uebertragungsfunktion des Vokaltraktes nicht nur durch eine Reihenschaltung von Formantnetzwerken, sondern auch durch eine Parallelschaltung

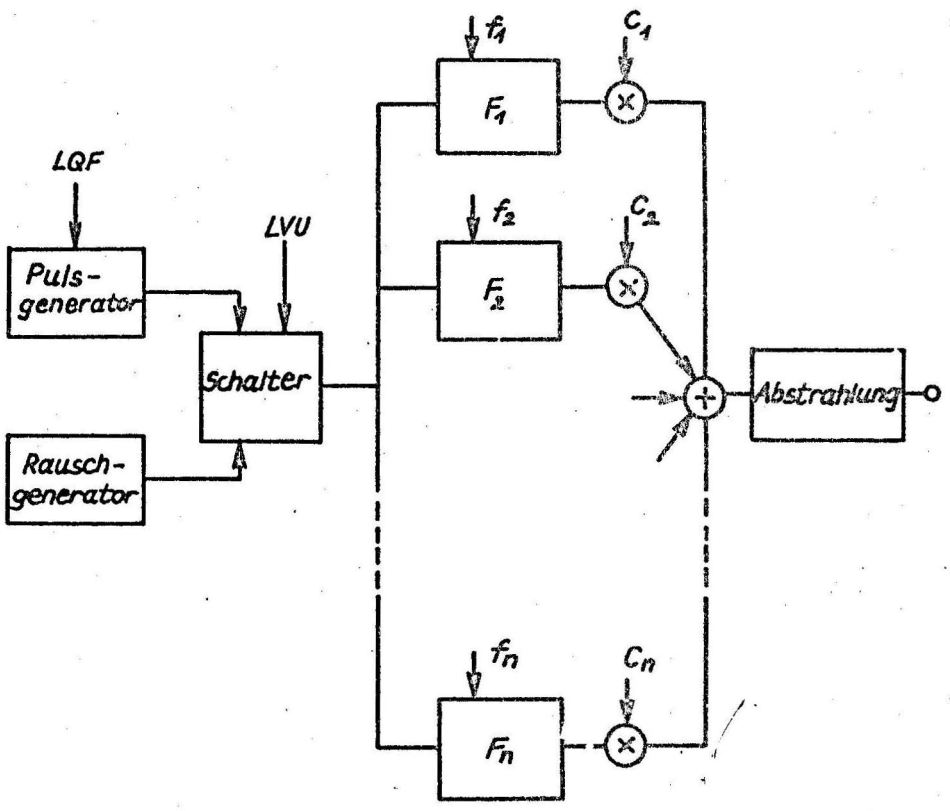


Abb.35, Formantsynthetisator als Parallelschaltung

darstellen laesst. Die Schaltung eines Formantvocoders in Parallelausfuehrung zeigt die Abb.35.
Ein grosser Nachteil der Parallelschaltung ist der, dass, wie auch aus Gl.(44) hervorgeht, zu jedem Formantglied eine eigene Amplitudenkontrolle C_n und eine Nullstelle $(s-s_{zn})$ hinzugefuegt werden muss.

Die Lagen der Nullstellen sind nicht konstant, sondern haengen von der jeweiligen Formantfrequenz ab. Die zusaetzlichen Koeffizienten C_n erhoehen die Bitrate, die zur Speicherung der Sprache aufgewendet werden muss. Da bei der Auswahl des Formantvocoders aus den anderen moeglichen Vocodertypen gerade auf den hohen Sprachkompressionsfaktor groessten Wert gelegt wurde, verzichtete der Verfasser auf diese Realisierungsmoeglichkeit, obwohl die Parallelausfuehrung auch Vorteile aufweist.

Ein Vorteil tritt dort auf, wo die Formanten durch digitale Rechenschaltungen mit endlicher Registerlaenge dargestellt werden. In diesen Faellen addiert sich das Quantisierungsrauschen bei den einzelnen Formanten, waehrend sich das Rauschen in einer Reihenschaltung der Formantnetzwerke multiplikativ ausbreitet.

Ein anderer Vorteil der Parallelschaltung tritt bei Verwendung von Formantnetzwerken auf, die nur einen kleinen Amplitudenbereich aufgrund eines geringen Signal- Rausch-Verhaeltnisses verarbeiten koennen, da bei der Parallelschaltung ein wesentlich kleinerer Dynamikbereich als bei der Reihenschaltung durchfahren wird und dadurch die Amplituden besser in Kohtrolle gehalten werden koennen.

Die genannten Nachteile der Parallelausfuehrung sind nach der Meinung des Verfassers schwerwiegender als die Vorteile, so dass sich der Verfasser fuer die serielle Anordnung der Formantglieder entschied.

5.2 Korrektur der hoeheren Pole

Nach Gl.(12) wird der Vokaltrakt bei der Erzeugung stimmhafter Laute durch eine unendlich grosse Anzahl von Formanten beschrieben, deren Frequenzlage sich naeherungsweise aus Gl.(13) zu

$$f = \pm \frac{(2n-1)\cdot c}{4l}$$

ergab.

Die Frequenzwerte berechneten sich dabei als alle ungeraden Vielfachen von 500 Hz, wie:

500 Hz, 1500 Hz, 2500 Hz, 3500 Hz, 4500 Hz, 5500 Hz, ...

Ein grosser Teil der Sprachinformation liegt, wie oben bereits erwaehnt wurde, in der Aenderung der Frequenzlagen der ersten drei Formanten. Man darf deshalb die hoeheren Formanten nicht einfach weglassen, da sonst das Spektrum verfaelscht wird. In Abb.36 ist die Uebertragungsfunktion des Vokaltraktes dargestellt, wenn man sie, wie das in Kap 2.2 geschah, durch ein einseitig geschlossenes Rohr konstanten Querschnitts approximiert. Darunter ist das Spektrum eines

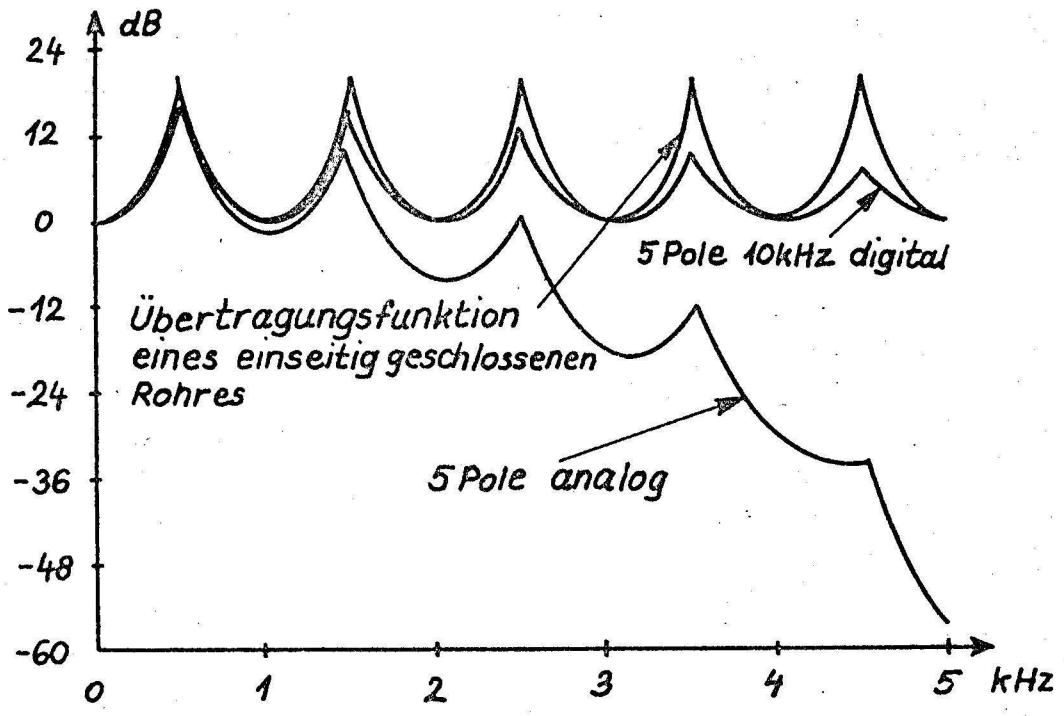


Abb.36, Anhebung der Spektren bei der Korrektur der hoeheren Pole

Formantsynthetisators bestehend aus 5 Formanten dargestellt, der in Analogtechnik aufgebaut worden ist. Die Spektralanteile des dritten, vierten und fuenften Formanten, die im Frequenzbereich unter 5 kHz liegen, fallen durch das Fehlen der hoeheren Formanten in ihrer Amplitude stark ab. Es muss deshalb eine Kompensation fuer die hoeheren Formanten,

bzw. ihren Einfluss auf den Frequenzbereich unter 5 kHz, vorgesehen werden. Das Gesagte gilt jedoch nur fuer die Serienausfuehrung eines Formantsynthetisators, da bei der Parallelausfuehrung sich die Spektren der einzelnen Formantglieder gegeneinander weniger beeinflussen.

Bei einem digitalen Formantvocoder ist die Korrektur der hoeheren Pole recht einfach. Wie in Kap 4.2 bereits erwaehnt wurde, setzen sich die Spektren diskreter Zeitfunktionen periodisch im Abstand der Abtastfrequenz fort. Man kann jetzt eine unendliche Anzahl von Formanten dadurch erzeugen, dass man das Spektrum des Vokaltraktes im Bereich bis zur halben Abtastfrequenz als gerade Funktion darstellt. Dann erscheint das Spektrum durch den Faltungseffekt genau noch einmal im Bereich von der halben bis zur ganzen Abtastfrequenz. Durch die genannte Periodizitaet der Spektren setzt sich das Spektrum des Formantnetzwerks bis in alle Unendlichkeit fort. Die Abb.37 verdeutlicht die Lage der Formanten in der komplexen Ebene fuer den Fall, dass ein gerader

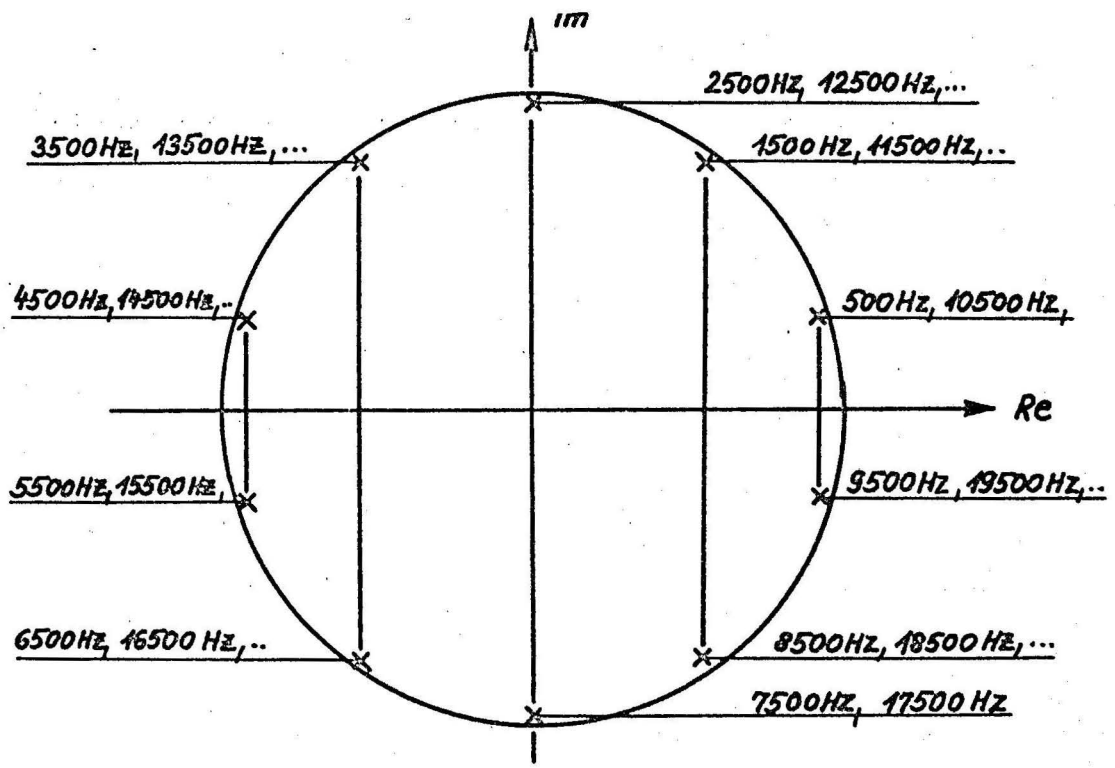


Abb.37, Frequenzlage der Formanten bei der Korrektur der hoeherer Formanten

Spektralverlauf bis zur halben Abtastfrequenz erzeugt werden soll, wenn die Abtastfrequenz 10 kHz betraegt. Aus Abb.37 geht hervor, dass die Formanten bei 500 Hz, 1500 Hz, 2500 Hz 3500 Hz und 4500 Hz liegen muessen, um den genannten Anforderungen zu entsprechen. Der zugehoerige Spektralverlauf ist in Abb.36 eingezeichnet.

Da die ersten drei Formanten, die in ihrer Frequenzlage variabel sein muessen, ohnehin im Bereich 500 Hz, 1500 Hz und 2500 Hz liegen, muessen zur Korrektur der hoeheren Pole bei einem digitalen Formantsynthetisator mit einer Tastfre-

quenz von 10 kHz lediglich zwei weitere Formanten mit den konstanten Frequenzwerten 3500 Hz und 4500 Hz hinzugefuegt werden.Die zugehoerigen Bandbreitewerte zu 175 Hz und 281 Hz wurden von RABINER /23/ uebernommen.

5.3 Bauelemente des Formantsynthetisators

Wie schon aus Abb.26 zu ersehen ist, werden die verschiedensten Bauelemente fuer die Syntheseschaltung eines Formantvocoders benoetigt. Dazu gehoeren
fuer die Quelle:

1. Pulsgenerator
2. Pulsformnetzwerk
3. Rauschgenerator
4. Umschalter

fuer das Formantfilter:

1. Formanten
2. Antiformanten

fuer die Abstrahlung:

Abstrahlungsnetzwerk

zum Angleich der Frequenzgaenge der idealisierten Bausteine an die wahren Verhaeltnisse im menschlichen Spracherzeugungssystem:

1. Tiefpaesse
2. Bandpaesse
3. Hochpaesse

Der Aufbau dieser Elemente soll im folgenden insoweit beschrieben werden, wie sie vom Verfasser simuliert wurden.

Rausch- und Pulsgenerator

Als Rauschgenerator wird ein rueckgekoppelter Schieberegisterrauschgenerator verwendet. Der Rauschgenerator liefert ein weisses Rauschen mit einer Gleichverteilung.

Der Pulsgenerator muss in seiner Pulsfolgefrequenz variabel gemacht werden. Er besteht aus einem einfachen Zaehler, der nach der einstellbaren Zeitdauer fuer eine Pitchperiode einen Puls liefert. Die Groesse des Pulses wird so berechnet, dass fuer eine mittlere Pitchfrequenz von 100 Hz der Effektivwert des Pulsgenerators einschliesslich des nachgeschalteten Pulsformnetzwerkes mit dem des Rauschgenerators uebereinstimmt. Auf diese Weise sollen der gleiche Lautstaerkeeindruck bei Puls- und Rauschgenerator sichergestellt werden.

Formant und Antiformant

Nach Gl.(6) ist ein Formant ein konjugiert komplexes Polpaar der Form:

$$H_p(s) = \frac{s_p \cdot s_p^*}{(s-s_p)\cdot(s-s_p^*)} \quad (6)$$

Bei der Simulation eines Formantnetzwerkes auf dem Digitalrechner hat man es mit diskreten Systemen zu tun. In dem Fall wird ein einfacher Resonator durch die Differenzengleichung zweiter Ordnung beschrieben:

$$y(nT) = k_1 y(nT-T) + k_2 y(nT-2T) + x(nT) \quad (45)$$

Es gelten dabei die Anfangsbedingungen y(-T)=0 und y(-2T)=0. Wendet man die z- Transformation auf Gl.(45) an, erhaelt man

$$\left.\begin{aligned} H(z) = \frac{Y(z)}{X(z)} = \frac{1}{1-k_1 z^{-1} - k_2 z^{-2}} = \frac{z^2}{(z-z_1)\cdot(z-z_2)} \\ \text{mit} \quad k_1 = z_1 + z_2 \qquad \text{und} \quad k_2 = -z_1 \cdot z_2 \end{aligned}\right\} (46)$$

Aus Gl.(46) geht hervor, dass der Resonator fuer ein diskretes System in der z-Ebene ebenfalls zwei Pole aufweist. Fuer einen Formanten muessen die Pole, wie in Gl.(6), konjugiert komplex sein. Mit der Polfrequenz ω_p und der Daempfung σ_p ergeben sich die Pollagen:

$$\begin{aligned} z_1 &= \exp[-\sigma_p \cdot T + j\omega_p \cdot T] \\ z_2 &= \exp[-\sigma_p \cdot T - j\omega_p \cdot T] \end{aligned} \quad (47)$$

Setzt man Gl.(47) in Gl.(46) ein und normiert die Gleichung entsprechend Gl.(6) so, dass

$$|H(z)|_{z=1} = 1 \qquad \text{ist,}$$

ergibt sich:

$$H(z) = \frac{1 - 2e^{-\sigma_p T}\cos\omega_p T + e^{-2\sigma_p T}}{1 - 2e^{-\sigma_p T}\cos\omega_p T\, z^{-1} + e^{-2\sigma_p T} z^{-2}} \quad (48)$$

Daraus ergibt sich mit

$$C = 1 - 2e^{-\sigma_p T}\cos\omega_p T + e^{-2\sigma_p T}$$

die rekursive Filtergleichung fuer ein Formantnetzwerk zu:

$$y(nT) = C[2e^{-\sigma_p T}\cos\omega_p T \cdot y^*(nT-T) - e^{-2\sigma_p T} y^*(nT-2T) + x(nT)] \quad (49)$$

Die Gleichung laesst sich durch das folgende Blockschaltbild veranschaulichen:

$$\begin{aligned} y^*(nT) &= \frac{y(nT)}{C} \\ y^*(nT-T) &= \frac{y(nT-T)}{C} \\ y^*(nT-2T) &= \frac{y(nT-2T)}{C} \end{aligned}$$

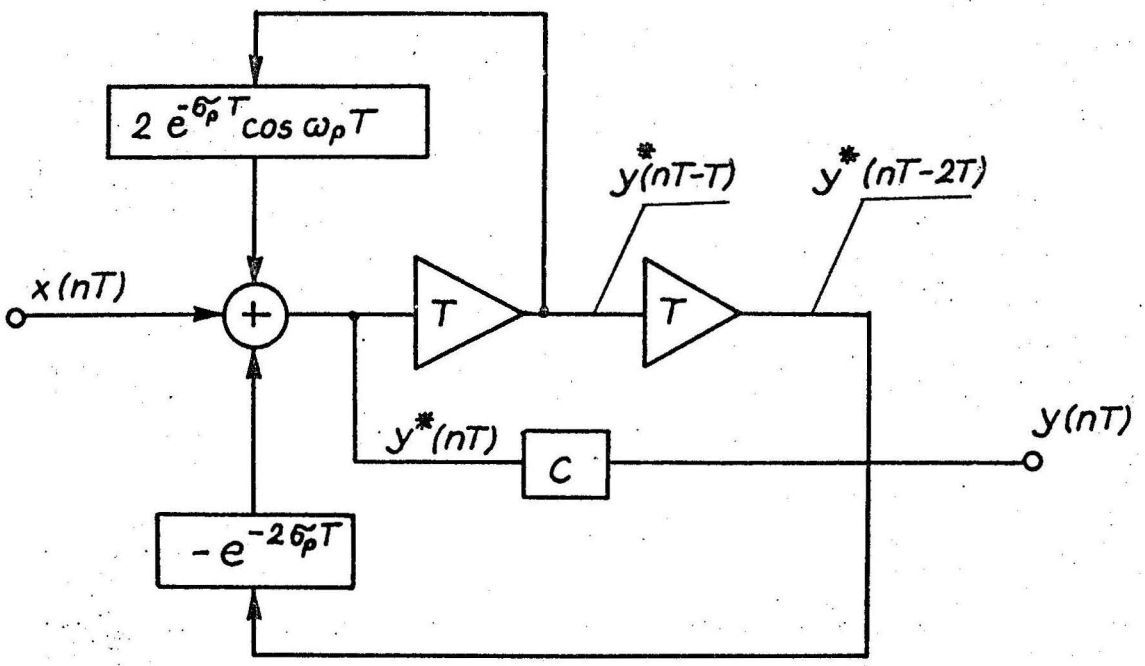


Abb.38, Blockschaltung eines Formantgliedes

Die Darstellung nach Abb.38 widerspricht insofern den anschaulichen Vorstellungen eines Formantfilters, als der Ausgang y(nT) und der Eingang x(nT) gleichzeitig erscheinen. Man koennte daher den Eingang um ein Samplingintervall verzoegern. Diese Betrachtungen sind jedoch fuer das Ergebnis der durchgefuehrten Simulation ohne jede Bedeutung.
Nach Gl.(7) ist ein Antiformant ein konjugiert komplexes Nullstellenpaar der Form:

$$H_z(s) = \frac{(s-s_z)(s-s_z^*)}{s_z \cdot s_z^*} \qquad (7)$$

Die Uebertragungsfunktion unterscheidet sich von der des Formanten nach Gl.(6) lediglich dadurch, dass Zaehler und Nenner vertauscht werden. Es liegt daher nahe, fuer die Ermittlung der z-Uebertragungsfunktion des Antiformanten eines diskreten Systems dasselbe zu tun. Dann ergibt sich aus der Gl.(48):

$$H_z(z) = \frac{1}{H_p(z)} = \frac{1-2e^{-\sigma_p T}\cos\omega_p T\, z^{-1} + e^{-2\sigma_p T} z^{-2}}{1-2e^{-\sigma_p T}\cos\omega_p T + e^{-2\sigma_p T}} \qquad (50)$$

Die zugehoerige rekursive Filtergleichung lautet:

$$y(nT) = [x(nT) - 2e^{-\sigma_p T}\cos\omega_p T \cdot x(nT-T) + e^{-2\sigma_p T} x(nT-2T)]/C \qquad (51)$$

Die Gl.(51) wird durch das Blockschaltbild in Abb.39 veranschaulicht.

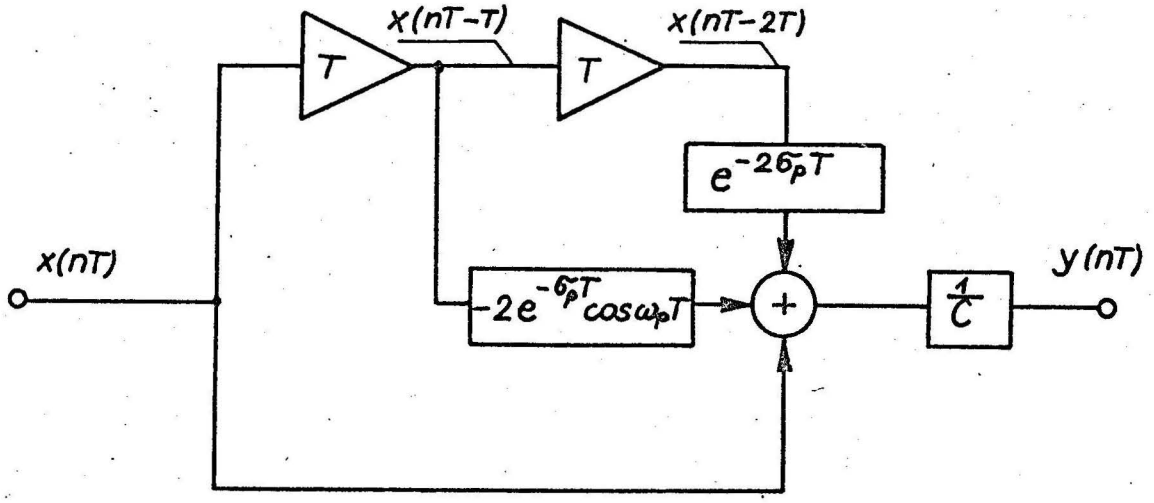


Abb.39, Blockschaltbild eines Antiformantgliedes

Pulsformnetzwerk

Nach FLANAGAN (/2/ S.44) faellt das Amplitudenspektrum der Pulsfolge an der Glottis mit 12 dB/Okt ab. Dieser Effekt erklaert sich auch aus dem dreieckfoermigen Zeitverlauf der Schnelle an der Glottis. Da ein Abfall von 12 dB/Okt einer Filterung mit einem zweipoligen Filter entspricht, liegt es nahe, ein Formantfilter fuer diesen Zweck zu verwenden. Da am Ausgang des Filters, das durch Pulse angeregt wird, ein dreieckfoermiger Verlauf ohne Nulldurchgaenge entstehen soll, muss das Filter stark gedaempft werden. Es wurde in dem Zusammenhang eine Poldaempfung von der doppelten Polfrequenz vorgesehen. Die Polfrequenz des Pulsformnetzwerkes und damit die zeitliche Dauer des Dreieckpulses wurde mit dem Parameter der Pitchfrequenz gekoppelt.

Es wurden experimentell der folgende Zusammenhang zwischen der Pitchfrequenz und sowohl der Polfrequenz als auch der Daempfung ermittelt:

$$\omega_p = \frac{3\pi}{2 \cdot T_p \cdot T} \qquad T_p = \text{Laenge der Pitchperiode}$$

$$\sigma_p = 2\omega_p$$

Den zeitlichen Verlauf am Ausgang des Pulsformnetzwerkes bei Anregung durch einen Einheitspuls zeigt die Abb.40.

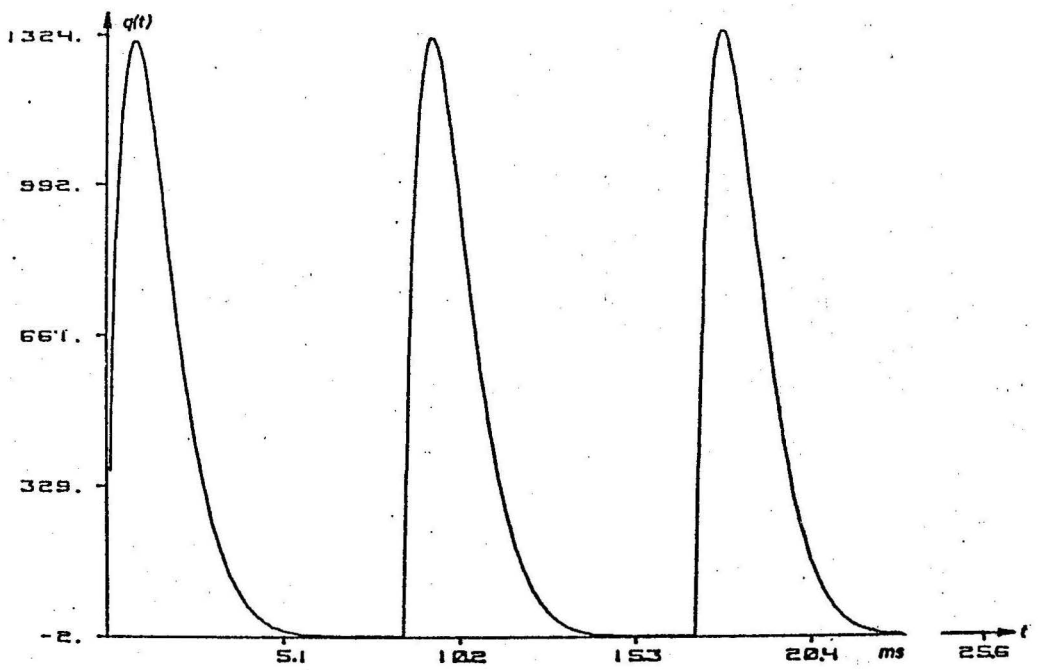


Abb.40, Impulsantwort des Pulsformnetzwerkes

Bandpass, Hochpass, Tiefpass

Bandpaesse mit verschiedenen Bandbreiten und Mittenfrequenzen, wie Hoch- und Tiefpaesse mit variablen Grenzfrequenzen wurden in Frequency- Sampling- Technique erstellt.

Auf die Frequency- Sampling- Technique wurde bereits in Kap 3.3 hingewiesen. In Abb.41 sind noch einmal die Bausteine eines solchen Filters dargestellt (/21/ S.85).
Die Abb.41a zeigt das Blockschaltbild eines Kammfilters, die Abb.41b die eines Elementarresonators, und die Abb.41c zeigt die Zusammenschaltung des Kammfilters mit 7 Elementarresonatoren zu einem Bandpass.

Die Abb.42 verdeutlicht die Technik noch einmal anhand der Lagen der Nullstellen des Kammfilters und der Pollagen der Resonatoren in der z- Ebene fuer einen Bandpass. Die Kreise auf dem Einheitskreis stellen die Lagen der Nullstellen dar. Sie haben voneinander den Abstand

$$\omega = \frac{2\pi}{m \cdot T}$$

wobei m die Anzahl der Nullstellen ist. Die Kreuze stellen die Pole der Resonatoren dar. Der Bereich, den die nebeneinanderliegenden Pole ueberstreichen, ist ein Mass fuer die Bandbreite des Bandpasses. Im Falle eines Filteraufbaus nach Abb.41c ist die Bandbreite des Bandpasses

$$\omega_b = (n-3) \cdot \frac{2\pi}{m \cdot T} \quad (52)$$

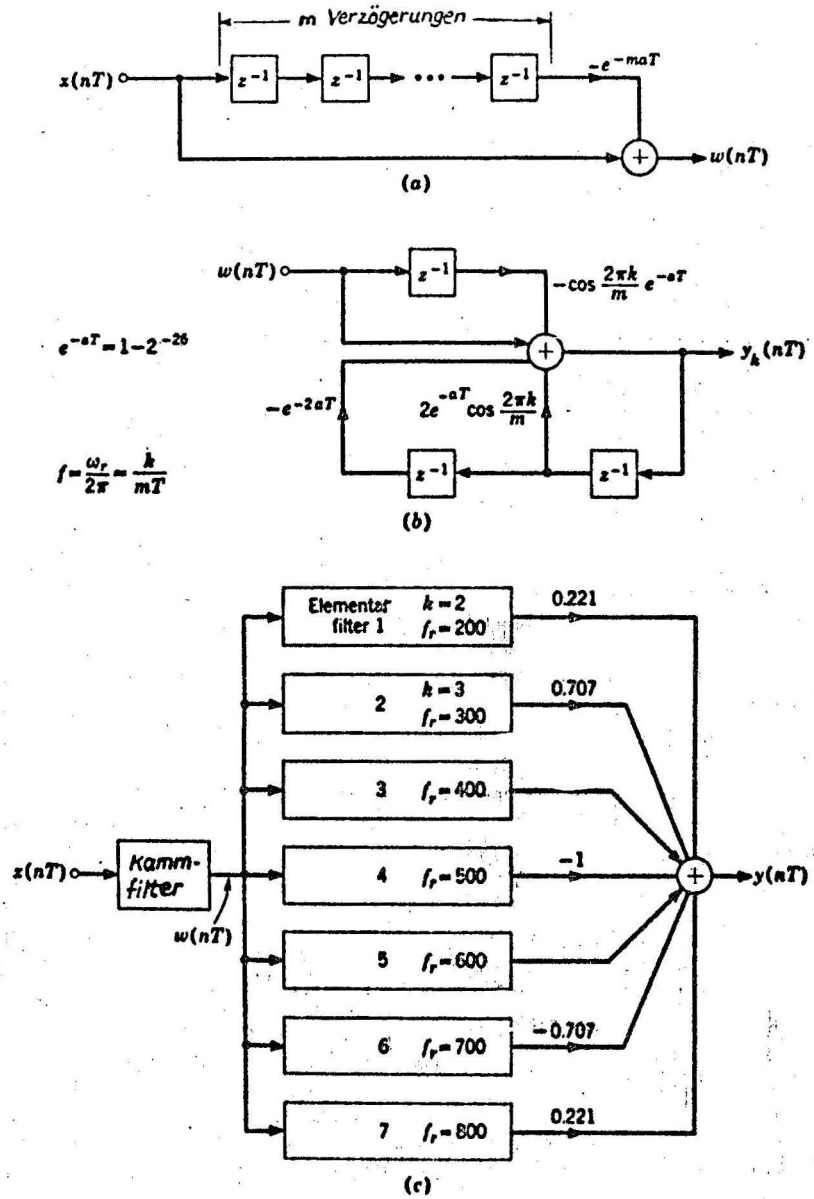


Abb.41, Aufbau eines digitalen Filters in Frequency- Sampling- Technique

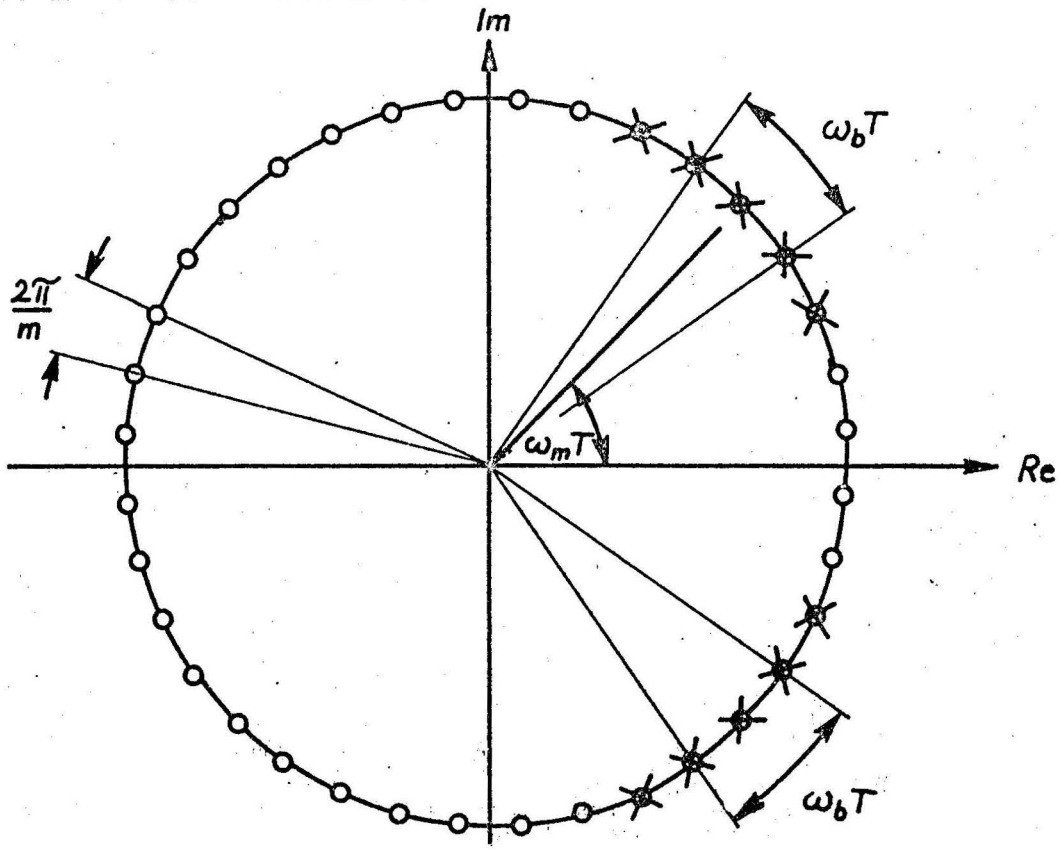


Abb.42, Pol- und Nullstellenlagen eines Filters in Frequency- Sampling- Technique in der z-Ebene (Bandpass)

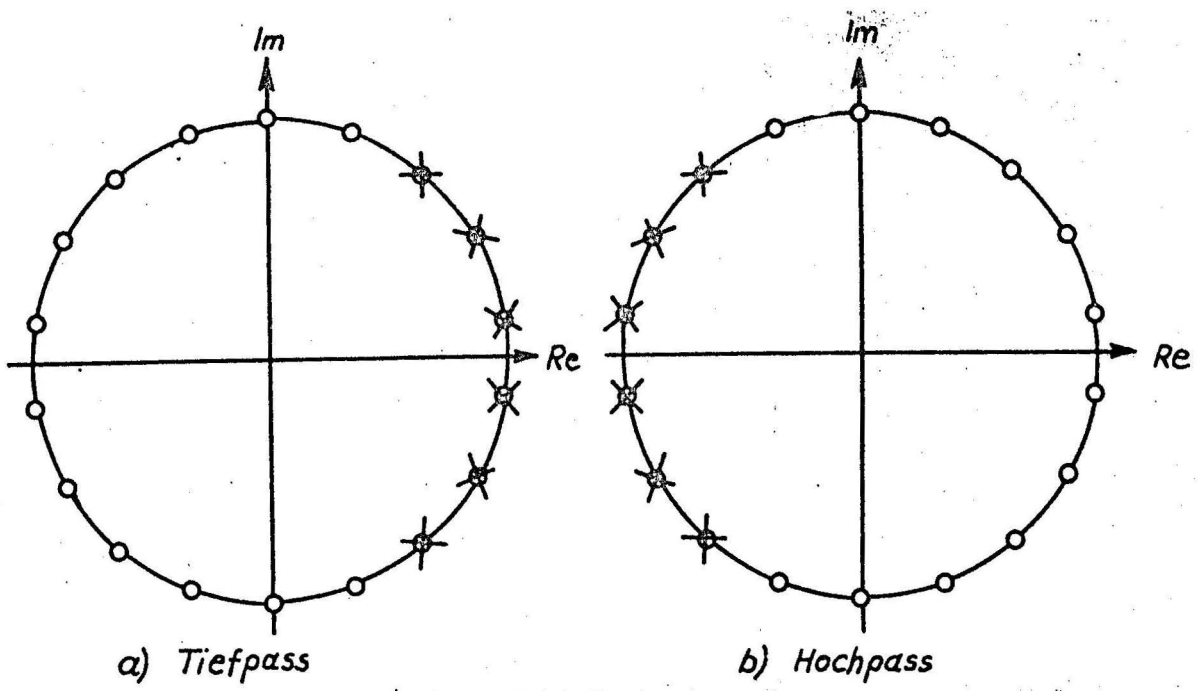


Abb.43, Tiefpass und Hochpass

wobei n die Anzahl der nebeneinanderliegenden Resonatoren ist. Die Zahl 3, die von n zu subtrahieren ist, ergibt sich aus der Groesse der Bewertungskoeffizienten (0.221, 0.707) in Abb.41c, die einen Abfall des Frequenzgangs auf den Faktor 0.221 bzw. 0.707 bei der entsprechenden Polfrequenz verursachen, und bei denen die Polfrequenzen daher ausserhalb des Durchlassbereiches liegen.

Die Mittenfrequenz des Bandpasses ergibt sich aus der Lage der mittleren Resonatoren. Ordnet man die n Resonatoren nach Abb.43a an, erhaelt man einen Tiefpass. Die Grenzfrequenz berechnet sich zu:

$$\omega_g = \left(\frac{n}{2} - \frac{3}{2}\right) \frac{2\pi}{m \cdot T} \qquad (53)$$

Bei der Anordnung nach Abb.43b erhaelt man einen Hochpass mit der Grenzfrequenz:

$$\omega_g = \left(\frac{m}{2} - \frac{n}{2} + \frac{3}{2}\right) \cdot \frac{2\pi}{m \cdot T} \qquad (54)$$

Aus Gl.(53) und Gl.(54) ist ersichtlich, dass ein Tiefpass einem Bandpass nach Gl.(52) entspricht mit $\omega_m = 0$ und $\omega_g = \omega_b/2$ und ein Hochpass einem Bandpass mit $\omega_m = \pi/T$ und $\omega_g = \omega_b/2$. Die Flankensteilheit der Filter ist eine Funktion der Zahl n der zum Aufbau des Filters verwendeten Resonatoren. Eine Erhoehung der Flankensteilheit bei gleicher Grenzfrequenz oder Bandbreite laesst sich dadurch erreichen, dass die Anzahl m der Nullstellen auf dem Einheitskreis vergroessert und entsprechend mehr Resonatoren zur Abdeckung der Nullstellen verwendet werden.

Fuer den praktischen Aufbau der Filter sei hier noch ein Hinweis gegeben (/21/ S.83): Durch die Quantisierungsfehler ist es nicht moeglich, dass die Pole stets die Nullstellen genau abdecken. Deswegen werden sowohl Pole als auch Nullstellen innerhalb des Einheitskreises gelegt, um auf jeden Fall Instabilitaeten zu vermeiden. GOLD und RADER geben dabei einen Radius

$$e^{-aT} = 1 - 2^{-26}$$

fuer die Lage der Pole und Nullstellen an.

Schneller Tiefpass

Eines der am haeufigsten verwendeten Bauelemente bei der Sprachverarbeitung sind Tiefpaesse. Der Aufbau eines Tiefpasses in Frequency- Sampling- Technique hat zwei grosse Nachteile:

1. hoher Speicherplatzbedarf
2. hoher Rechenzeitbedarf

Ein wesentlicher Grund fuer den hohen Speicherplatzbedarf ist die Verwendung des Kammfilters. Es besteht aus einem Schieberegister mit so vielen Gliedern, wie Nullstellen auf

dem Einheitskreis auftreten. Der hohe Rechenzeitbedarf ergibt sich teils ebenfalls aus der Verwendung des Kammfilters, dessen Schieberegister bei jedem Rechenschritt weitergeschoben werden muss und teils aus der grossen Anzahl von Resonatoren, bei denen jeder drei Verzoegerungsglieder enthaelt, wie aus Abb.41 hervorgeht. Ein sehr grosser Vorteil des Filters ist, dass sehr leicht beliebig grosse Flankensteilheiten erzeugt werden koennen, natuerlich wieder auf Kosten des Speicherplatzbedarfes und der Rechenzeit.

In vielen Faellen genuegt ein Tiefpass mit einer Flankensteilheit von 18 dB/Okt und es waere wuenschenswert, fuer diesen Fall einen Tiefpass zu entwerfen, der eine wesentlich kuerzere Rechenzeit benoetigt. Als 'schneller Tiefpass' wurde daher ein dreipoliger Tschebyscheff- Tiefpass verwendet. Den Aufbau des digitalen Filters zeigt die Abb.44. Die Ko-

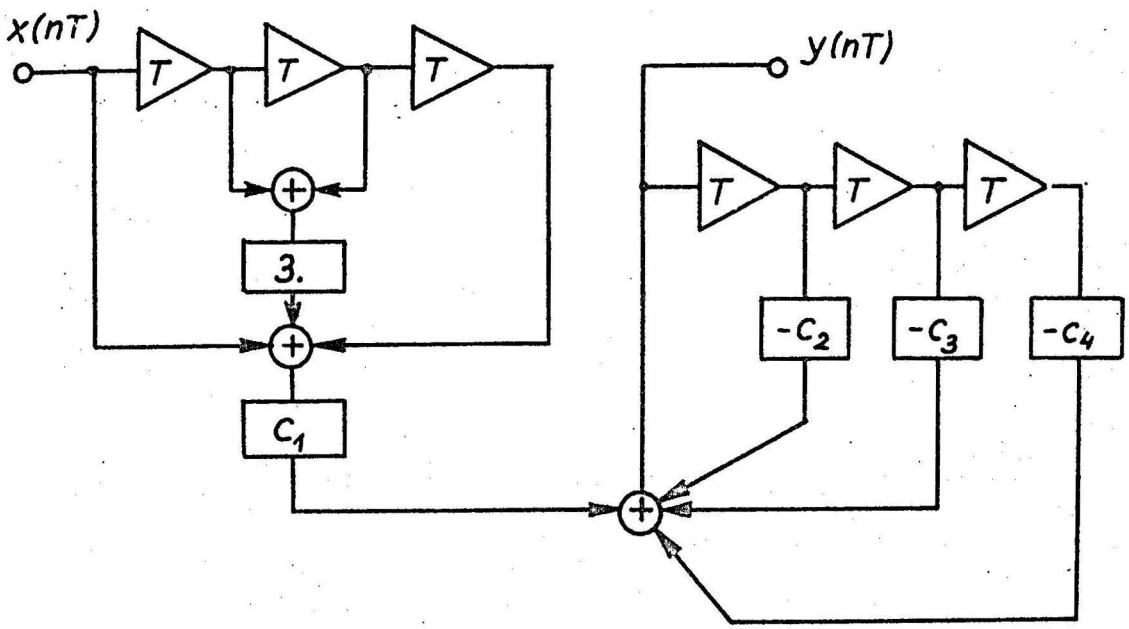


Abb.44, Blockschaltbild des Tschebyscheff- Tiefpass als digitales Filter

effizienten C_1, C_2, C_3 und C_4 berechnen sich aus der Grenzfrequenz und der Welligkeit im Durchlassbereich (/20/ S.31).

Abstrahlungsnetzwerk

Wie in Kap 2.2 bereits erwaehnt wurde, laesst sich die Abstrahlung durch ein Filter beruecksichtigen, das in der komplexen s-Ebene einen Pol auf dem negativen Teil der reellen Achse und eine Nullstelle im Ursprung aufweist. Als Poldaempfung wurde σ_p =3500 Hz gewaehlt. Es muss nun ein digitales Filter, und um Speicherplatz zu sparen, moeglichst ein rekursives Filter entwickelt werden, das den gleichen Fre-

quenzgang aufweisen soll. Zum Entwurf wurde das Verfahren nach DECZKY /34/ verwendet:

Voraussetzung fuer das Verfahren nach DECZKY ist, dass sich die Uebertragungsfunktion des gewuenschten Filters als Quotient zweier Funktionsverlaeufe darstellen laesst. Die Funktionsverlaeufe des Zaehlers und Nenners brauchen dabei nur punktweise vorzuliegen.

Beruecksichtigt man, dass das Spektrum einer diskreten Zeitfunktion sich periodisch fortsetzt und dass das Spektrum der abgetasteten, analogen Zeitfunktion in Regel- und Kehrlage um alle Vielfache der Abtastfrequenz erscheint, so folgt daraus, dass der fuer die Berechnung der z-Uebertragungsfunktion angesetzte Spektralverlauf bezueglich den Vielfachen der Abtastfrequenz stets eine gerade Funktion ist. Dann laesst sich auch das Quadrat vom Betrage des Spektrums als gerade Funktion darstellen. Nimmt man fuer Zaehler und Nenner getrennt eine Fourieranalyse vor, so ergibt sich:

$$|H(z)|^2_{z=e^{j\omega T}} = \frac{\sum_{k=0}^{N} a_k \cos k\omega T}{\sum_{l=0}^{M} b_l \cos l\omega T} \qquad (55)$$

Da das Spektrum eines digitalen Filters dem Verlauf des Betrages der z-Uebertragungsfunktion auf dem Einheitskreis in der z-Ebene entspricht, kann die Substitution $z=exp[j\omega T]$ erfolgen:

$$|H(z)|^2_{z=e^{j\omega T}} = \left. \frac{\sum_{k=0}^{N} a_k \frac{z^k+z^{-k}}{2}}{\sum_{l=0}^{M} b_l \frac{z^l+z^{-l}}{2}} \right|_{z=e^{j\omega T}} \qquad (56)$$

Aus Gl.(56) geht hervor, dass aufgrund gleicher Koeffizienten fuer die Glieder

$$z^k \text{ und } z^{-k} \text{ bzw. } z^l \text{ und } z^{-l}$$

zum Einheitskreis spiegelbildliche Nullstellen und Pole auftreten. Deshalb kann man Gl.(56) auch schreiben:

$$|H(z)|^2_{z=e^{j\omega T}} = \left. \frac{P(z) \cdot P(z^{-1})}{Q(z) \cdot Q(z^{-1})} \right|_{z=e^{j\omega T}} \qquad (57)$$

wobei P und Q Polynomen sind, die gerade von der Haelfte aller Wurzeln des Zaehler- und Nennerpolynomens von Gl.(56) gebildet werden. Durch Aufspaltung entsprechend Gl.(57) kann die gewuenschte z-Uebertragungsfunktion des Filters als

$$H(z) = \frac{P(z)}{Q(z)} \qquad (58)$$

gewonnen werden. Es ist dabei zu beachten, dass Q(z) gerade alle Pole innerhalb des Einheitskreises enthalten muss, damit das Filter stabil ist.

Zur Berechnung des Filters ist folgendes zu bemerken: Die Gl.(55) enthaelt Summen aus N bzw. M Gliedern, ebenso wie die gewonnene z-Uebertragungsfunktion nach Gl.(58) ein Zaehlerpolynomen N-ter und Nennerpolynomen M-ter Ordnung aufweist. Entscheidet man sich aber fuer die Ordnungszahl $n<N$ fuer das Zaehlerpolynomen der z-Uebertragungsfunktion,

benoetigt man nur n+1 Koeffizienten aus dem Zaehler der Gl.(55). Wird die Reihe aber nach n+1 Gliedern willkuerlich abgebrochen, fuehrt das i.a. zu einer schlechten Approximation des vorgegebenen Spektralverlaufes, wenn die folgenden Glieder der Fourierreihe nicht von vornherein vernachlaessigbar klein gegenueber den ersten n+1 waren. Man kann die Approximation dadurch verbessern, indem man die Reihe mit einer Fensterfunktion bewertet, die ein langsames Abklingen des Betrages der Fourierkoeffizienten bis einschliesslich des n+1-ten Gliedes erzwingt. Die restlichen Glieder liegen ausserhalb des Fensters und werden mit Null bewertet. Vom Verfasser wurde in dem Zusammenhang der positive Zweig der Gaussschen Glockenkurve als Fensterfunktion verwendet.

Die Abb.45 zeigt den vorgegebenen Spektralverlauf und das Spektrum des nach DECZKY berechneten rekursiven Filters bei einer vorgegebenen Ordnungszahl von 6 sowohl fuer das Zaehler- als auch fuer das Nennerpolynomen.

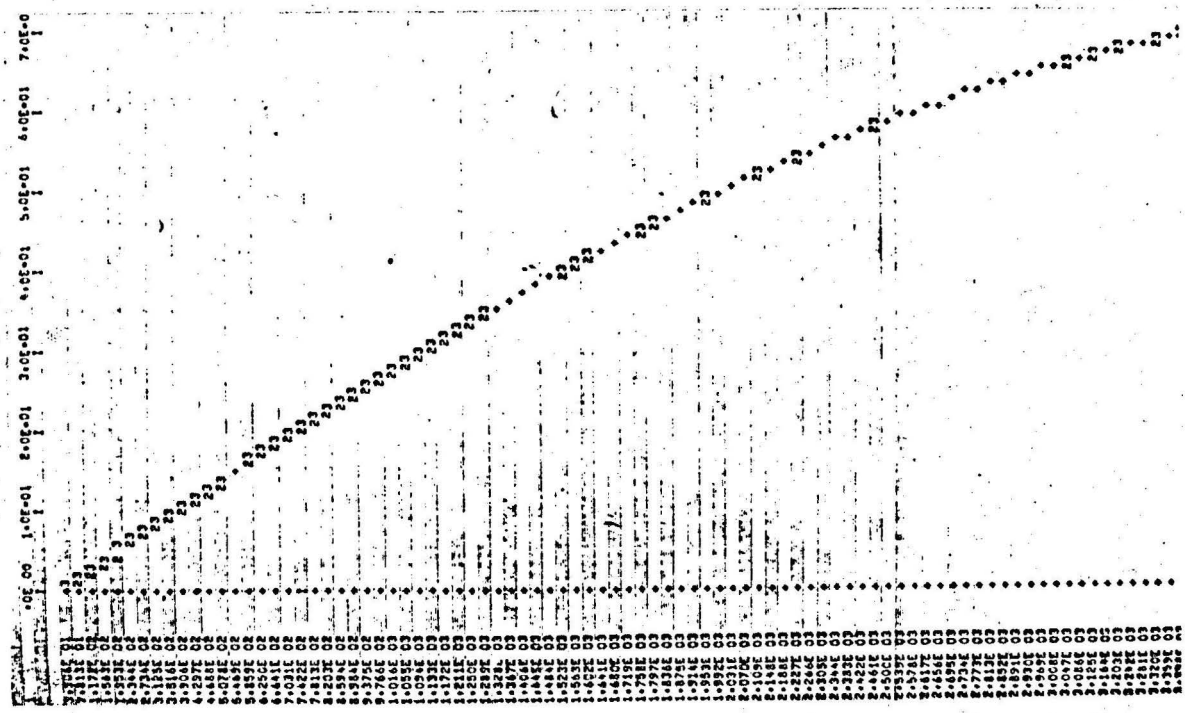


Abb.45, Vorgegebener und realisierter Spektralverlauf des Abstrahlungsnetzwerkes. Der gegebene Spektralverlauf wird durch '3' und der approximierte Zeitverlauf durch '2' dargestellt.

5.4 Programmsystem zur digitalen Simulation

Um die verschiedenen Kombinationsmoeglichkeiten von Formanten und Antiformanten studieren, sowie die Notwendigkeit der Filterung einzelner Teilsignale mit Hochpaessen, Tiefpaessen und Bandpaessen ueberpruefen zu koennen, wurde ein Simulationssystem fuer einen Formantvocoder entwickelt. Die Bausteine dieses Simulationssystems wurden in ihrer Eigenschaft als digitale Filter bereits in 5.3 beschrieben.

Bezeichnung	Vorbereitungsphase	Typ	Rechenphase	Typ
Pulsgenerator	VPULS(A)	S	RPULS(B)	S
Rauschgenerator	VRAUSCH(A)	S	RRAUSCH(B)	S
Umschalter			MIX(A,B,C)	S
Pulsformnetzwerk	VSHAPE	S	SHAPE(X)	F
Tiefpass	VTP1(,) VTP2(,) .	S S	TP1(X) TP2(X) .	F F
Bandpass	VBP1(,) VBP2(,) .	S S	BP1(X) BP2(X) .	F F
Formanten	VFORM1 VFORM2 .	S S	FORM1(X) FORM2(X) .	F F
Netzwerk zur Korrektur der hoeheren Pole	VFORM4 VFORM5	S S	FORM4(X) FORM5(X)	F F
Antiformanten	VZERO1 VZERO2 .	S S	ZERO1(X) ZERO2(X) .	F F
Abstrahlungsnetzwerk	VRAD	S	RAD(X)	F

Tabelle 7, Bauelemente des Simulationsprogramms

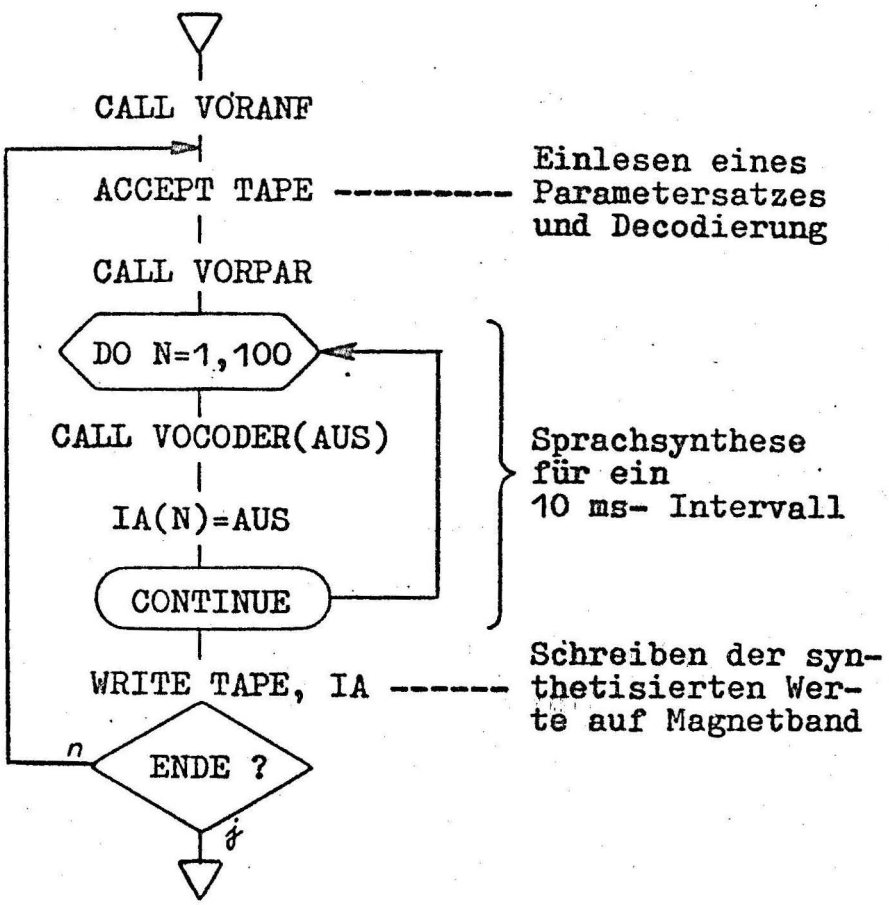
Die digitalen Filter werden in dem Simulationsprogramm jeweils durch zwei Unterprogramme dargestellt, von denen das

eine die sog. Vorbereitungsphase und das andere die Rechenphase darstellt. Bei den Bauelementen handelt es sich, wie oben bereits erwaehnt wurde, um Speicherplatz zu sparen, um rekursive Filter. In der Vorbereitungsphase werden die Koeffizienten der rekursiven Filter aus den gewuenschten Filterparametern, wie Grenzfrequenz, Bandbreite, Daempfung, Polfrequenz usw. berechnet. Die Koeffizienten werden ueber einen COMMON- Bereich dem Unterprogramm zur Verfuegung gestellt, das die Rechenphase darstellt. In der Rechenphase wird dann die synthetische Sprachzeitfunktion erzeugt.

Die Tabelle 7 gibt einen Ueberblick ueber die verwendeten Bauelemente und ihre Realisierung durch jeweils zwei Unterprogramme. Der Typ des Unterprogramms ist mit S fuer SUBROUTINE und F fuer FUNCTION gekennzeichnet.

Die Bauelemente sind alle unabhaengig voneinander und koennen daher in jeder beliebigen Anordnung verwendet werden. Abb.46 zeigt den Aufbau des Syntheseprogramms. Es enthaelt im Wesentlichen die drei SUBROUTINE-Unterprogramme VORANF, VORPAR, und VOCODER(AUS). Saemtliche Parameter werden ueber den COMMON- Bereich uebergeben.

Das Unterprogramm VORANF enthaelt saemtliche Vorbereitungsphase, die insgesamt nur einmal abgearbeitet werden. VORPAR enthaelt im Gegensatz dazu alle Vorbereitungsphasen,



\bb.46, Aufbau des Syntheseprogramms

die nach jeder Parameteruebergabe neu aufgerufen werden

muessen.

Saemtliche Rechenphasen der Syntheseschaltung sind im Unterprogramm VOCODER(AUS) vereinigt, wobei die Variable AUS den berechneten Syntheseswert darstellt.

Das Programm ist sehr flexibel aufgebaut und gestattet leicht das Hinzufuegen oder die Aenderung von Vocoderbauelementen. Das soll hier einmal an dem Beispiel eines Formantvocoders, der dem in Abb.26 aehnelt, demonstriert werden. Zu diesem Zweck ist der Syntheseteil des Formantvocoders noch einmal in Abb.47 unter Angabe aller Parameter dargestellt.

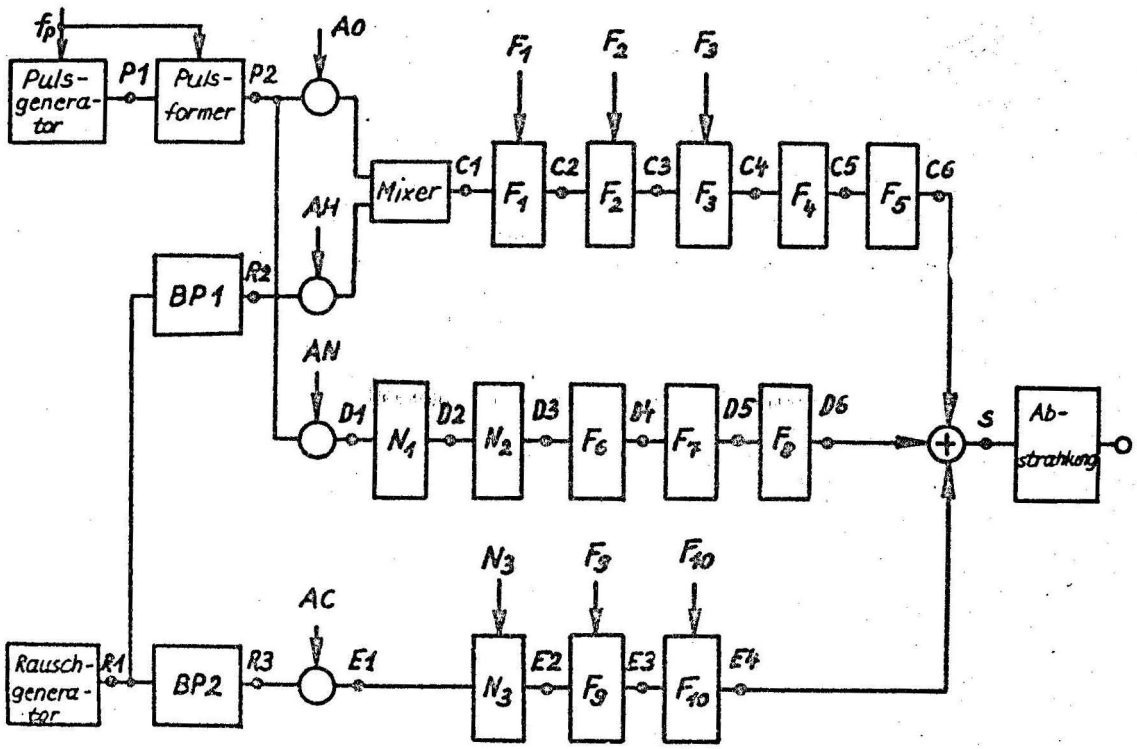


Abb.47, Beispiel fuer den Syntheseteil eines Formantvocoders

In der folgenden Liste sind die zugehoerigen Unterprogramme VORANF, VORPAR und VOCODER(AUS) dargestellt.

```
-----------------------------------------

      SUBROUTINE VORANF
C     =================
C
C-----GENERATOR
      CALL VPULS(26.)
      CALL VRAUSCH(1.)
      CALL VBP1(2000.,2500.)
      CALL VBP2(2500.,2300.)
```

```
C
C-----FORMANTFILTER FUER VOKALE
      CALL VFORM4
      CALL VFORM5
C
C-----FORMANTFILTER FUER NASALE
      CALL VZERO1
      CALL VZERO2
      CALL VFORM6
      CALL VFORM7
      CALL VFORM8
C
C-----ABSTRAHLUNG
      CALL VRAD
      RETURN
      END

-----------------------------------------

      SUBROUTINE VORPAR
C     =================
C
C-----PULSFORMNETZWERK
      CALL VSHAPE
C
C-----FORMANTFILTER FUER VOKALE
      CALL VFORM1
      CALL VFORM2
      CALL VFORM3
C
C-----FORMANTFILTER FUER FRICATIVE
      CALL VZERO3
      CALL VFORM9
      CALL VFORM10
      RETURN
      END

-----------------------------------------

      SUBROUTINE VOCODER(AUS)
C     =======================
C
C-----GENERATOR
      CALL RPULS(P1)
      P2=SHAPE(P1)
      CALL RRAUSCH(R1)
      R2=BP1(R1)
      R3=BP2(R1)
C
C-----FORMANTFILTER FUER VOKALE
      CALL MIX(C1,A0*P2,AH*R2)
      C2=FORM1(C1)
      C3=FORM2(C2)
      C4=FORM3(C3)
      C5=FORM4(C4)
      C6=FORM5(C5)
```

```
C
C-----FORMANTFILTER FUER NASALE
      D1=AN*P2
      D2=ZERO1(D1)
      D3=ZERO2(D2)
      D4=FORM6(D3)
      D5=FORM7(D4)
      D6=FORM8(D5)
C
C-----FORMANTFILTER FUER FRICATIVE
      E1=AC*R3
      E2=ZERO3(E1)
      E3=FORM9(E2)
      E4=FORM10(E3)
C
C-----SUMMIERUNG
      S=C6+D6+E4
      AUS=RAD(S)
      RETURN
      END
```

5.5 Synthese auf dem Digitalrechner

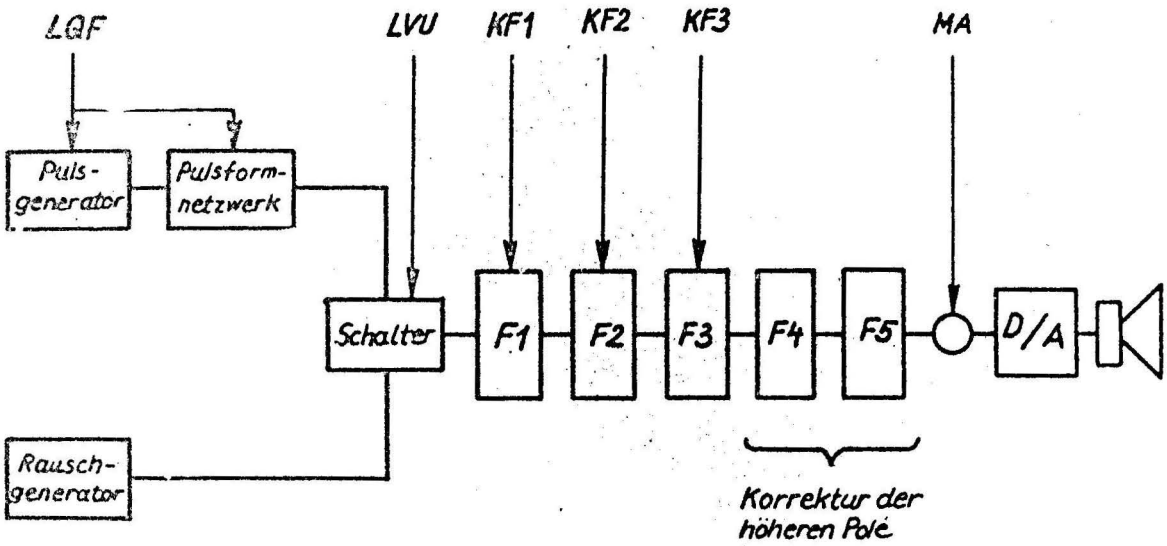


Abb.48, Verwendeter Formantsynthetisator

In der Abb.48 ist das Blockschaltbild des Formantsynthetisators dargestellt, mit dem saemtliche Sprachsynthesen durchgefuehrt worden sind.

Die Quelle besteht aus einem Pulsgenerator, dem ein Pulsformnetzwerk, wie es unter 5.3 beschrieben wurde, nachgeschaltet ist. Zur Erzeugung stimmloser Laute steht der oben genannte Rauschgenerator zur Verfuegung, der an seinem Ausgang ein weisses Rauschen mit einer Gleichverteilung liefert. Die Rauschquelle und der Pulsgenerator lassen sich ueber einen Schalter abwechselnd auf das nachfolgende Formantfilter schalten. Auf eine Filterung des Rauschsignals wurde gaenzlich verzichtet, da sich hierdurch keinerlei erkennbarer Einfluss auf die Verstaendlichkeit der Sprache feststellen liess.

Das Formantfilter weist im Hinblick auf eine spaetere hardwaremaessige Realisierung des Synthetisators eine sehr einfache Struktur auf. Es besteht zur Erzeugung saemtlicher Laute nur aus einem einzigen Zweig, der wiederum nur aus Formantgliedern zusammengesetzt ist. Die ersten drei Formanten koennen in ihrer Frequenz variiert werden. Die zugehoerigen Daempfungswerte werden aus einer Tabelle entnommen. Die Tabelle wurde nach Angaben von FLANAGAN (/2/ S.152) berechnet. Der vierte und fuenfte Formant liegen in ihren Frequenz- und Bandbreitewerten fest und dienen zur Korrektur der hoeheren Pole.

Auf die Beruecksichtigung der Abstrahlung konnte ver-

zichtet werden, da die Sprache ohnehin durch einen Lautsprecher am Ausgang abgestrahlt wird.

Am Ausgang des Formantfilters befindet sich die Amplitudenkontrolle MA. Sie erzwingt einen Amplitudenverlauf, der dem der analysierten Sprachzeitfunktion entspricht.

Es werden insgesamt 6 Steuerparameter benoetigt. Dabei handelt es sich um:

LQF --- Laenge der Pitchperiode
LVU --- Stimmhaft- Stimmlosigkeit
KF1 --- erste Formantfrequenz
KF2 --- zweite Formantfrequenz
KF3 --- dritte Formantfrequenz
MA --- Amplitudenkontrolle

Der Parameter LVU kann folgende Werte annehmen:

LVU=0 fuer stimmlose Laute
LVU=1 fuer stimmhafte Laute
LVU=2 fuer den Zwischenraum zwischen Lauten

Um abrupte Uebergaenge zwischen Gebieten LVU≠2 und LVU=2 zu vermeiden, wurde im Syntheseprogramm vorgesehen, dass bei LVU=2 die Quelle vom Formantfilter abgetrennt wird, so dass die Filter zwar nicht mehr angeregt, aber langsam ausschwingen koennen. Der Aufbau des Simulationsprogramms entspricht im Prinzip dem, das in Abb.46 durch ein Flussdiagramm veranschaulicht worden ist.

5.6 Synthese auf dem Hybridrechner (/15/,/17/)

Die Rechenzeit zur Erzeugung von 1 sek Sprache auf dem Digitalrechner betraegt selbst bei einer so einfachen Syntheseschaltung, wie sie in Abb.45 dargestellt wurde, ca 5 min. Um die Rechenzeit zu verkuerzen, wurde deshalb ein Formantsynthetisator als analoge Rechenschaltung auf dem zur Verfuegung stehenden Analogrechner RA 770 programmiert. Die analoge Rechenschaltung wird ueber das Koppelwerk HKW 900 vom Digitalrechner CAE 90-40 gesteuert.

Der Aufbau des Formantsynthetisators ist in Abb.49a dargestellt. Das Formantfilter enthaelt zwei getrennte Zweige fuer die Erzeugung stimmhafter und stimmloser Laute. Das Formantfilter zur Erzeugung stimmhafter Laute besteht aus drei Formanten, und der Zweig zur Erzeugung stimmloser Laute besteht aus zwei Formanten und einem Antiformanten.

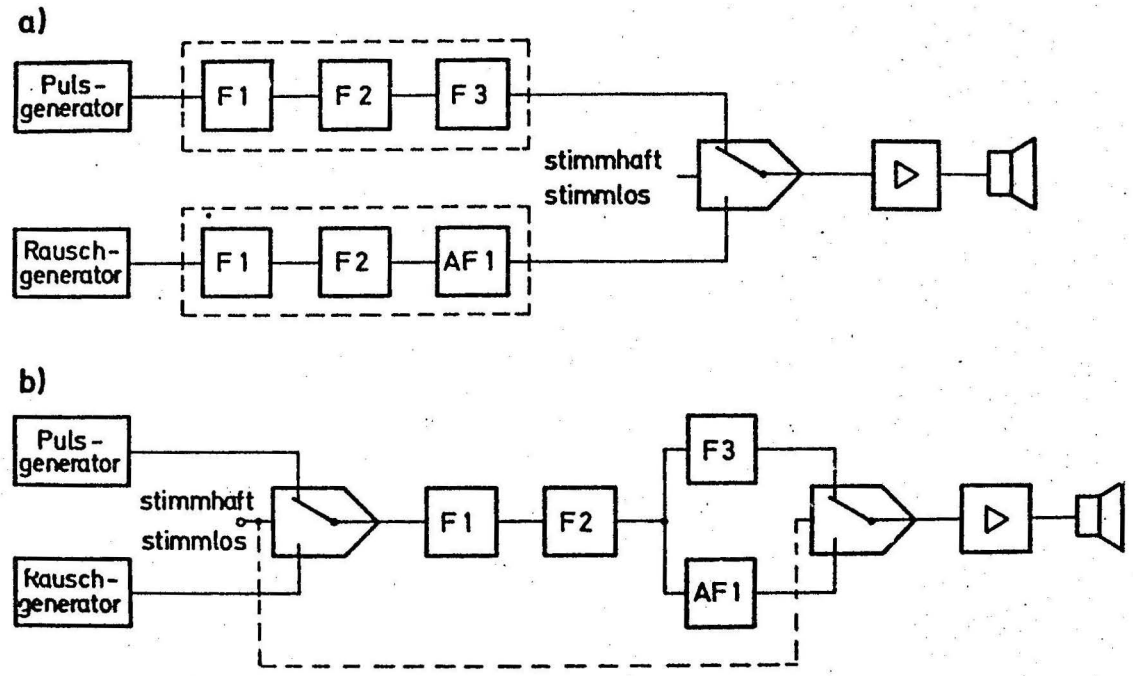
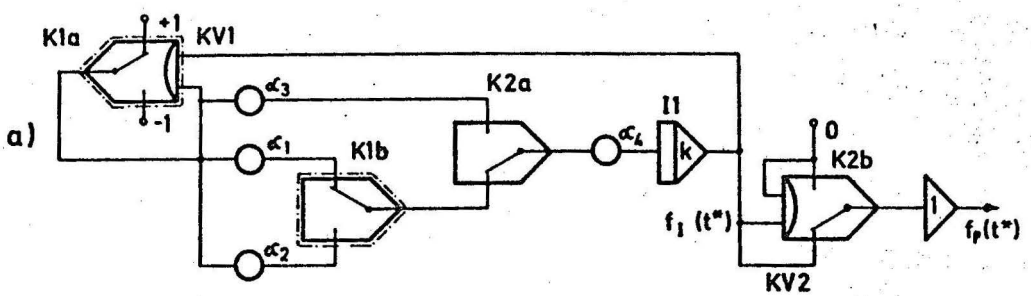


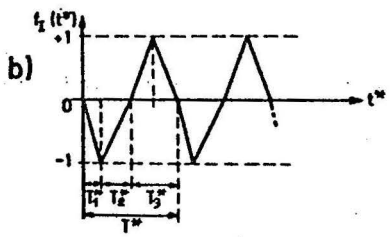
Abb.49, Konfiguration des hybriden Formantvocoders

Um die Anzahl der verwendeten Rechenelemente so gering wie moeglich zu halten, wurden der erste und zweite Formant, die beiden Zweigen gemeinsam sind, zusammengefasst. Puls- und Rauschgenerator, sowie der dritte Formant und die Nullstelle wurden umschaltbar ausgefuehrt, so dass sich fuer die Ana-

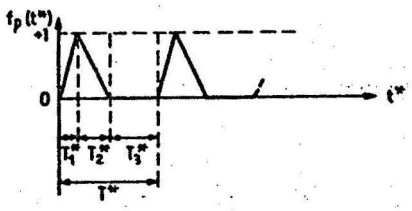
logschaltung ein Aufbau nach Abb.49b ergibt.



Prinzipschaltbild des Pulsgenerators



Spannungsverlauf am Integrier-ausgang



Spannungsverlauf am Ausgang des Pulsgenerators

Abb.50, Pulsgenerator

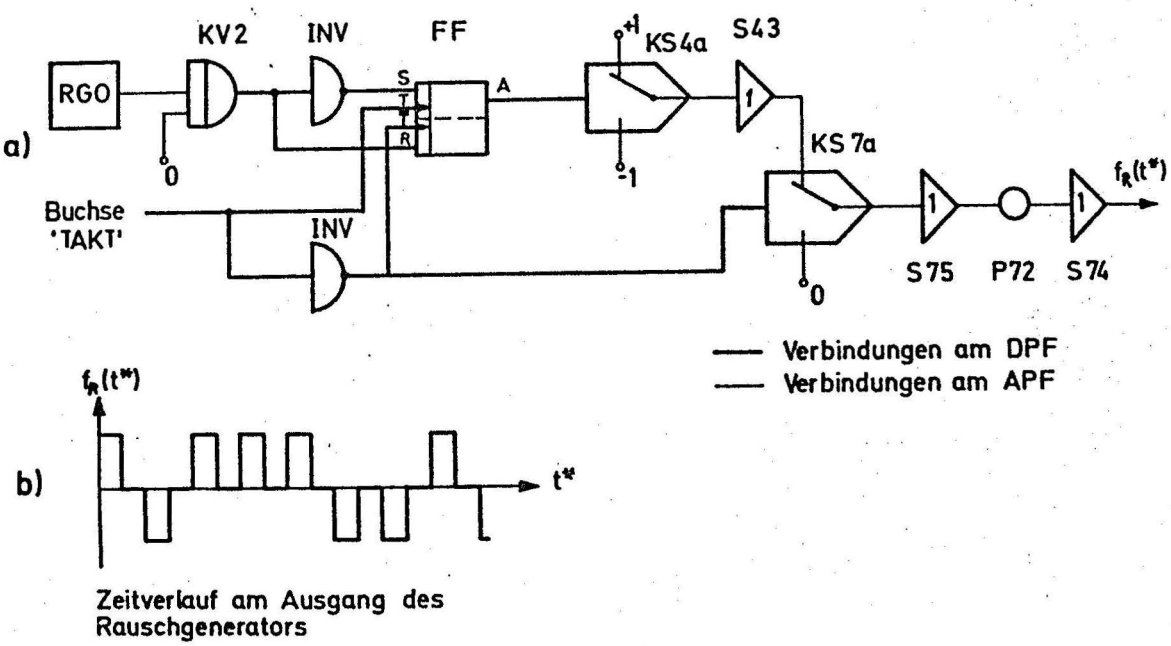
Pulsgenerator

Die Abb.50 zeigt den prinzipiellen Verlauf des Pulsgenerators. In Abb.50b sind der Zeitverlauf des Ausgangssignales und der Ausgang des Integrierers I1, $f_I(t^*)$ zu sehen. Der Komparatorverstaerker KV1 steuert die Schalter K1a und K1b und der Komparatorverstaerker KV2 die Schalter K2a und K2b. Die Potentiometerstellung α_4 sei zunaechst konstant angenommen. Die negative Flanke am Integriererausgang wird dann in ihrem Vorzeichen durch den Komparator K1a und in ihrer Steigung durch die Potentiometerstellung α_1 bestimmt. Bei $f_I(t^*)=-1$ schaltet der Komparatorverstaerker KV1 die Schalter K1a und K1b. K1a ruft einen Vorzeichenwechsel der Integrationsrichtung hervor, wobei die Steigung des Signals $f_I(t^*)$ jetzt durch α_2 bestimmt wird. Wenn $f_I(t^*)$ positiv wird, schaltet der Komparatorverstaerker KV2 und der Betrag der Steigung wird bis zum Ende der Periode durch α_3 bestimmt. Das Ausgangssignal $f_p(t^*)$ ist waehrend des positiven Verlaufs von $f_I(t^*)$ zu Null gesetzt. Der Vorteil der Schaltung ist der, dass die Anstiegszeit T1, die Abfallzeit T2 und die Pausenzeit T3 unabhaengig voneinander eingestellt werden koennen. Die Summe der 3 Zeiten ergibt die Periodendauer T=T1+T2+T3.

Durch Variation von α_4 laesst sich eine beliebige Pulsfolgefrequenz einstellen, ohne dass sich die Proportionen des Dreieckspulses aendern.

Rauschgenerator

Es steht ein Rauschgenerator zur Verfuegung, dem eine Gaussverteilung der Amplituden zugrundeliegt und dessen Leistungsdichtespektrum eine Grenzfrequenz von 4.5 kHz besitzt. Dieser Rauschgenerator laesst sich aber nicht als Erregerquelle zur Erzeugung stimmloser Laute verwenden, da er das nachfolgende Formantnetzwerk schlecht aussteuert. Als Rauschgenerator wurde deshalb ein Rechteckpulsgenerator verwendet, bei dem die Zufallsinformation im Vorzeichen der Pulse liegt. Abb.51a zeigt die Rechenschaltung und Abb.51b einen Ausschnitt aus der erzeugten Zeitfunktion.



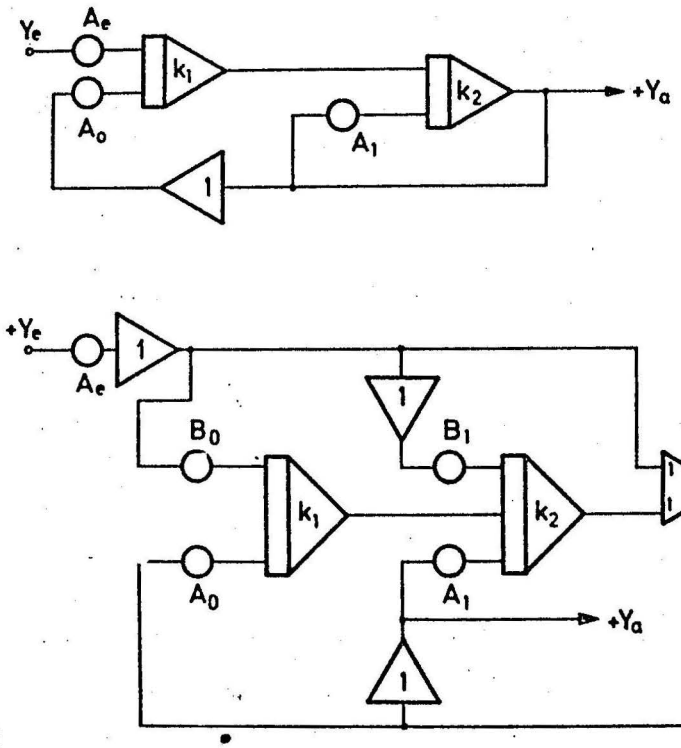
Zeitverlauf am Ausgang des Rauschgenerators

Abb.51, Rauschgenerator

Das Vorzeichen am Ausgang des Rauschgenerators RGO wird zu aeuqidistanten Zeitpunkten abgefragt und im Flipflop FF auf dem Digitalzusatz des Analogrechners gespeichert. Das Flipflop steuert den Schalter KS4a. Der Ausgang von KS4a, der +1 oder -1 betragen kann, wird ueber den Schalter KS7a mit einem festen Takt moduliert und gibt das gewuenschte Ausgangssignal entsprechend Abb.51b.

Formantnetzwerk

Das Filter, das dem Puls-Rauschgenerator nachgeschaltet ist, setzt sich aus Formanten und Antiformanten zusammen. Die Rechenschaltung, die einen Formanten, also ein Polpaar in der komplexen Ebene darstellt, wird in Abb.52 gezeigt.



a) Rechenschaltung für einen Formanten

$$G(S)=\frac{Y_a(S)}{Y_e(S)}=\frac{Sp\cdot Sp^*}{(S-Sp)\cdot(S-Sp^*)}$$

b) Rechenschaltung für einen Antiformanten

$$G(S)=\frac{Y_a(S)}{Y_e(S)}=\frac{(S-S_N)\cdot(S-S_N^*)}{S_N\cdot S_N^*}$$

Abb.52, Formant- und Antiformantglied

Y_e stellt das Eingangssignal und Y_a das Ausgangssignal dar. Das Potentiometer A_0 dient zur Einstellung der Polfrequenz und A_1 zur Einstellung der Daempfungswerte. Das Potentiometer A_e wird zur Aussteuerung der Formantschaltung benutzt.

Der Antiformant, der in der komplexen Ebene einem Nullstellenpaar entspricht, hat nach Gl.(7) die Uebertragungsfunktion

$$G(s)=\frac{(s-s_z)\cdot(s-s_z^*)}{s_z\cdot s_z^*} \qquad \begin{matrix} s_z=-\sigma_z+j\omega_z \\ s_z^*=-\sigma_z-j\omega_z \end{matrix} \qquad (7)$$

Ein durch diese Uebertragungsfunktion beschriebenes System kann jedoch physikalisch nicht realisiert werden, da der Grad des Zaehlers groesser ist als der Grad des Nenners. Es muss daher ein konjugiert komplexes Polpaar hinzugefuegt

werden, dessen Polfrequenzen von den Frequenzwerten der Nullstellen moeglichst weit entfernt liegen, damit im Amplitudenspektrum der Einfluss des Nullstellenpaares erst bei hohen Frequenzen kompensiert wird. Die Gleichung der Uebertragungsfunktion lautet dann:

$$G(s) = \frac{s_p \cdot s_p^*}{(s-s_p)(s-s_p^*)} \quad \frac{(s-s_z)(s-s_z^*)}{s_z \cdot s_z^*} \qquad (59)$$

Die zugehoerige Rechenschaltung zeigt die Abb.52b. Darin bedeutet A_e das Eingangspotentiometer der gesamten Schaltung. A_0 bestimmt die Polfrequenz und A_1 die Poldaempfung. Mit B_0 wird die Nullstellenfrequenz und mit B_1 die Daempfung eingestellt.

Steuerung der Analogschaltung

Die Steuerparameter wurden in Abb.53 noch einmal herausgestellt. Es bedeuten LVU die Stimmhaft- Stimmlosigkeit und LQF die Pulsfolgezeit des Pulsgenerators. KF1, KF2, KF3 sind die Formantfrequenzen und SIG1, SIG2, SIG3 die zugehoerigen Daempfungswerte. Ausserdem wird der Parameter MVOC benoetigt, der ein Mass fuer die Ausgangsamplituden darstellt.

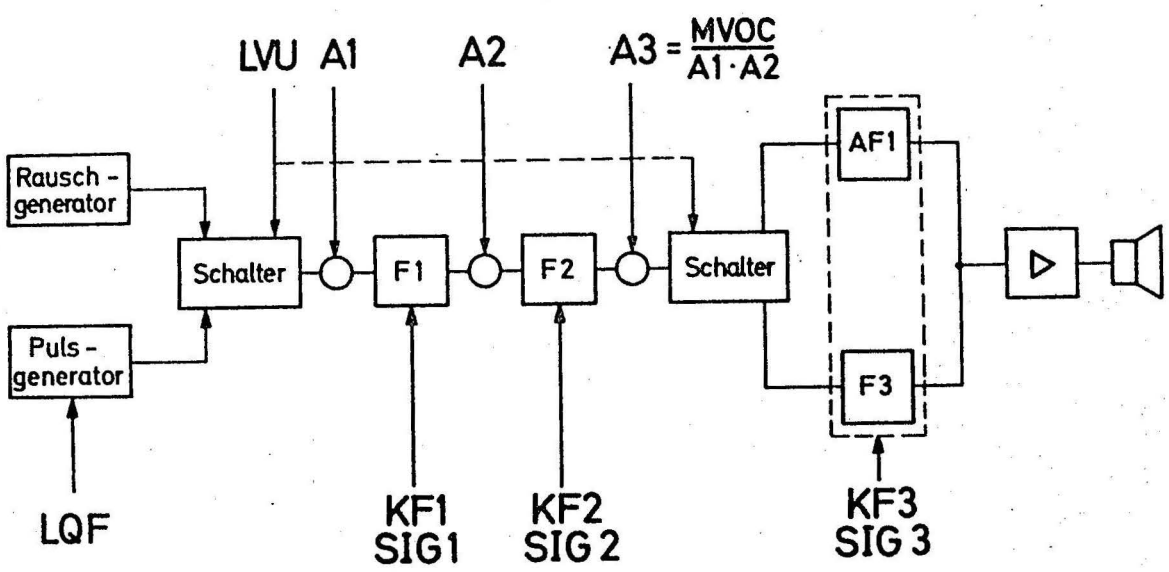


Abb.53, Steuerparameter des Formantvocoders

Fuer die Analyse wurde die Sprache in 10 ms- Intervalle auf-

gespalten und fuer jedes 10 ms- Intervall ein Parametersatz bestehend aus den 9 oben genannten Parametern erstellt. Die Sprachsynthese wird in der vierzigfachen Realzeit durchgefuehrt und daher muss alle 400 ms ein neuer Parametersatz an die Analogschaltung uebergeben werden. Die Parameteruebergabe erfolgt ueber multiplizierende Digital-Analog-Umsetzer. Lediglich der Parameter LVU, der hier nur die Zustaende LVU=0 fuer stimmlos und LVU=1 fuer stimmhaft annehmen kann, wird ueber eine Controlline uebergeben.

Die Gesamtschaltung besteht aus einer Reihenschaltung mehrerer voneinander entkoppelter Uebertragungssysteme. Bei fest eingestellten Potentiometern A1, A2, A3 besteht die Gefahr, dass einzelne Elemente entweder uebersteuert oder unzureichend ausgesteuert werden. Um das zu verhindern, wurden die drei Eingangspotentiometer A1, A2, A3 ebenfalls durch multiplizierende Digital-Analog-Umsetzer ersetzt und vom Digitalrechner gesteuert. Die Potentiometer A1 und A2 werden so eingestellt, dass am Ausgang des ersten und zweiten Formanten gerade Vollaussteuerung herrscht. Die Einstellung des dritten Potentiometers A3 berechnet sich dann aus

$$A3=MVOC/(A1\cdot A2)$$

wobei MVOC das Mass fuer die Ausgangsamplitude darstellt.

Ergebnis

Die erzielten Resultate sind in ihrer Qualitaet etwas geringer als die, die mit denselben Steuerparametern auf einem rein digitalen Formantvocoder nach 5.5 erzielt wurden. Ein Grund dafuer ist der, dass bei dem hybriden Formantvocoder auf eine Korrektur fuer die hoeheren Formanten verzichtet werden musste, da nicht genuegend viele Digital-Analog-Umsetzer zur Aussteuerungsregelung zur Verfuegung standen.

6. Analyseteil des Formantvocoders

6.1 Trennung gefalteter Komponenten

Der Druckverlauf bzw. das Amplitudenspektrum des Druckverlaufes im Schallfeld eines sprechenden Menschen berechnet sich nach Gl.(18) und Gl.(19). Wenn man in diesen Gleichungen den Einfluss des Abstrahlungsfaktors, der zeitlich konstant ist, vernachlaessigt, ergibt sich die vereinfachte Darstellung nach Gl.(60) und Gl.(61).

$$p(t) = q(t) * h(t) \qquad (60)$$

$$P(s) = Q(s) \cdot H(s) \qquad (61)$$

Zur Sprachanalyse steht eine Zeitfunktion zur Verfuegung, deren Verlauf proportional dem Druckverlauf nach Gl.(60) ist. Diese Zeitfunktion muss zunaechst in ihre Komponenten q(t) bzw. Q(s) und h(t) bzw. H(s) aufgespalten werden. Aus q(t) bzw. Q(s) werden die Parameter ermittelt, die die Quelle beschreiben. Dieser Teil der Analyse wird im folgenden Pitchbestimmung genannt.

Aus h(t) bzw. H(s) werden die Parameter ermittelt, die die Eigenschaften des Formantfilters beschreiben. Da es sich dabei im wesentlichen um die Lage der Formanten in der komplexen Ebene handelt, wird dieser Teil der Analyse im folgenden mit Formantbestimmung bezeichnet.
Die Auftrennung von p(t) in q(t) und h(t) bzw. die Aufspaltung von P(s) in Q(s) und H(s) ist nur moeglich, da die Komponenten unterschiedlichen Frequenzbereichen angehoeren. Die Pitchfrequenz liegt bei einem erwachsenen, maennlichen Sprecher im Bereich von 80 Hz bis 200 Hz, die Lagen der ersten drei Formanten dagegen im Bereich von 200 Hz bis 3 kHz.

Es gibt im wesentlichen zwei Verfahren zur Abspaltung von q(t) bzw. h(t) aus p(t).

1. Inverse Filterung
2. Homomorphe Filterung

Inverse Filterung

Die inverse Filterung setzt voraus, dass der Verlauf einer der beiden miteinander gefalteten Zeitfunktionen bekannt ist. Ist in dem vorliegenden Fall entsprechend Gl.(60) und Gl.(61) h(t) bzw. H(s) bekannt, so kann man durch lineare Filterung mit einem Filter der Uebertragungsfunktion

$$H^*(s) = \frac{1}{H(s)} \qquad (62)$$

den Anteil q(t) aus p(t) herausfiltern, wie Gl.(63) und die Gl.(64) zeigen:

$$Y(s) = \frac{1}{H(s)} \left[H(s) \cdot Q(s) \right] = Q(s) \qquad (63)$$

bzw.

$$q(t) = \mathcal{L}^{-1}[Q(s)] = \mathcal{L}^{-1} \left\{ \frac{1}{H(s)} \left[H(s) \cdot Q(s) \right] \right\} \qquad (64)$$

Homomorphe Filterung

Die Trennung oder Unterdrueckung von Signalkomponenten mit linearen Filtern ist nur dann optimal moeglich, wenn die betreffenden Komponenten additiv miteinander verknuepft sind. Da die Komponenten der Sprachzeitfunktion durch Faltung, also nichtlinear miteinander verknuepft sind, scheidet eine Auftrennung der Teilkomponenten durch lineare Filterung praktisch aus. Von ALAN OPPENHEIM /35/ wurde unter dem Namen 'homomorphe Filterung' eine Theorie zur nichtlinearen Filterung entwickelt. Die Theorie bemueht sich, die nichtlineare Filterung auf die mit viel Erfolg verwendeten Methoden der linearen Filterung zurueckzufuehren. Die Abb.54 zeigt die kanonische Darstellung eines homomorphen Filters.

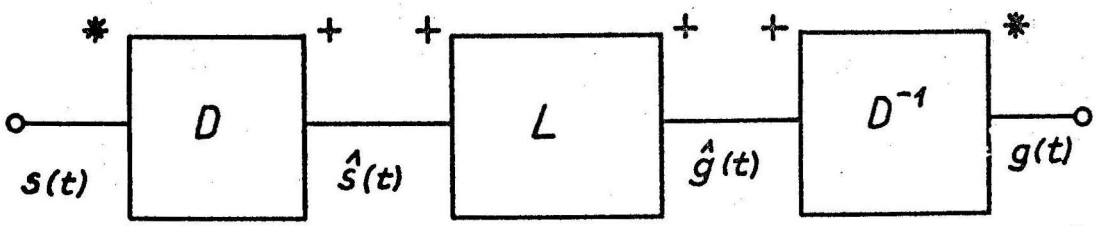


Abb.54, Homomorphes Filter

Wie aus Abb.54 hervorgeht, wird das Signal, das urspruenglich aus den miteinander gefalteten Komponenten $s_1(t)$ und $s_2(t)$ besteht, zunaechst durch ein homomorphes Filter mit der Eigenschaft D

$$D[s_1(t) * s_2(t)] = \hat{s}_1(t) + \hat{s}_2(t) \qquad (65)$$

in die additiv verknuepften Komponenten $\hat{s}_1(t)$ und $\hat{s}_2(t)$ zer-

legt. Dabei ist

$$\hat{s}_1(t) = D[s_1(t)] \qquad \hat{s}_2(t) = D[s_2(t)] \qquad (66)$$

Die additiv verknuepften Bestandteile $\hat{s}_1$ und $\hat{s}_2$ koennen linear gefiltert werden. Ein weiteres homomorphes Filter mit der inversen Eigenschaft zu D, also:

$$D^{-1}\{D(x(t))\} = x(t) \qquad (67)$$

verknuepft die linear gefilterten Bestandteile wieder durch Faltung.
Fuer den Anwendungsfall der Sprachverarbeitung, in dem die Phasenbeziehungen der Teilkomponenten ohne Interesse sind, kann das charakteristische System D durch die Einfuehrung des Cepstrums realisiert werden.

Das Cepstrum sei durch Gl.(68) eingefuehrt:

$$c_p[x(t)] = \mathcal{F}^{-1}\{\log|\mathcal{F}\{x(t)\}|^2\} \qquad (68)$$

Mit Gl.(68) ergibt sich Gl.(65) zu:

$$c_p[x_1(t) * x_2(t)] = c_p[x_1(t)] + c_p[x_2(t)] \qquad (69)$$

Das bedeutet, dass bei Anwendung von Gl.(69) auf Gl.(60) im Cepstrum der Zeitfunktion das Cepstrum der Quelle und das Cepstrum des Formantfilters additiv miteinander verknuepft vorliegen und aufgrund ihrer unterschiedlichen Frequenzbereiche voneinander getrennt werden koennen.

6.2 Pitchbestimmung

6.2.1 Pitchbestimmung aus dem Cepstrum

Das Cepstrum wurde bereits unter 6.1 nach Gl.(68) eingefuehrt. An dieser Stelle moechte der Verfasser darauf hinweisen, dass es verschiedene Definitionen fuer das Cepstrum gibt. Nach NOLL (/16/ S.300) ist das Cepstrum durch

$$c(\tau) = \left\{ \int_0^\infty |F(\omega)|^2 \cos(\omega\tau)\, d\omega \right\}^2 \quad (70)$$

definiert. Die unterschiedlichen Definitionen nach Gl.(70) und Gl.(68) haben keinen Einfluss auf die Pitchbestimmung aus dem Cepstrum. Der Verfasser haelt sich im folgenden immer an die Definition nach Gl.(68).

Wie bereits in 6.1 erwaehnt wurde, sind im Cepstrum des Sprachsignals das Cepstrum der Quelle und das Cepstrum des Vokaltraktes additiv miteinander verknuepft. Die Pitchfrequenz liegt im Bereich von ca 80 Hz bis 200 Hz. Daher liegt der Anteil der Quelle bei stimmhaften Lauten im Cepstrum im Quefrenzbereich von 5.0 ms bis 12.5 ms. Da die Formantfrequenzen im Bereich von 200 Hz bis 3 kHz liegen, nehmen sie im Cepstrum den Bereich von 0.3 ms bis 5.0 ms ein. Daraus geht hervor, dass die niederquefrenten Teile des Cepstrums von der Uebertragungsfunktion des Vokaltraktes und die hochquefrenten Anteile, zumindest bei stimmhaften Lauten, von der Quelle herruehren.

Die Abb.55a zeigt das Cepstrum eines stimmhaften Lautes und die Abb.55b das eines stimmlosen Lautes. Das Cepstrum

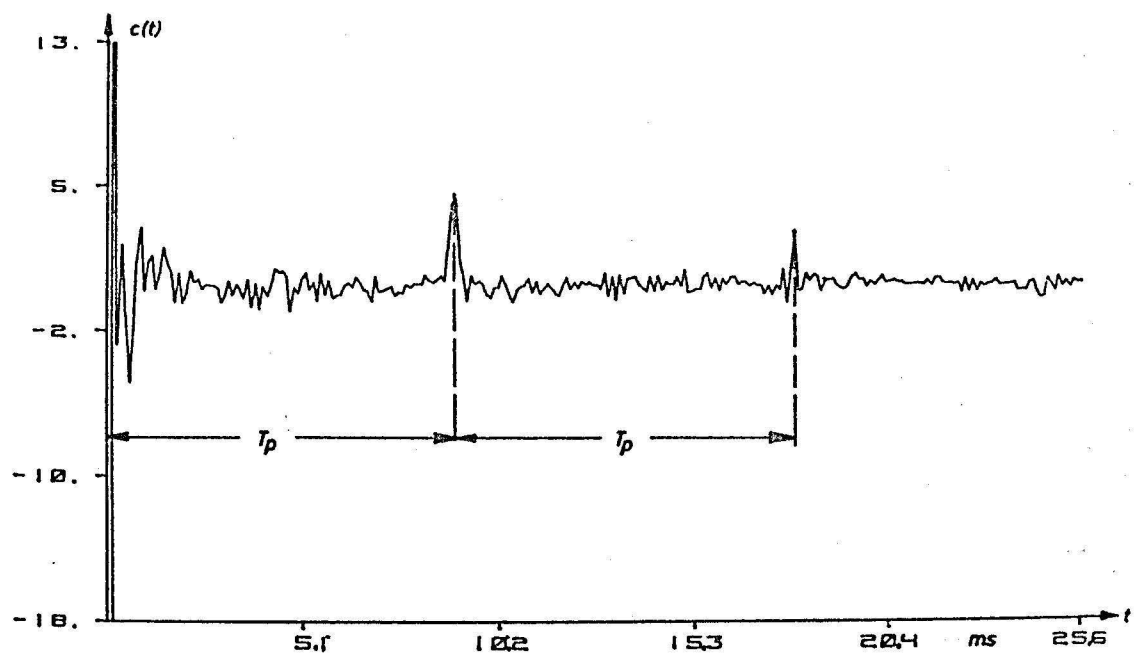


Abb.55a, Cepstrum eines stimmhaften Lautes

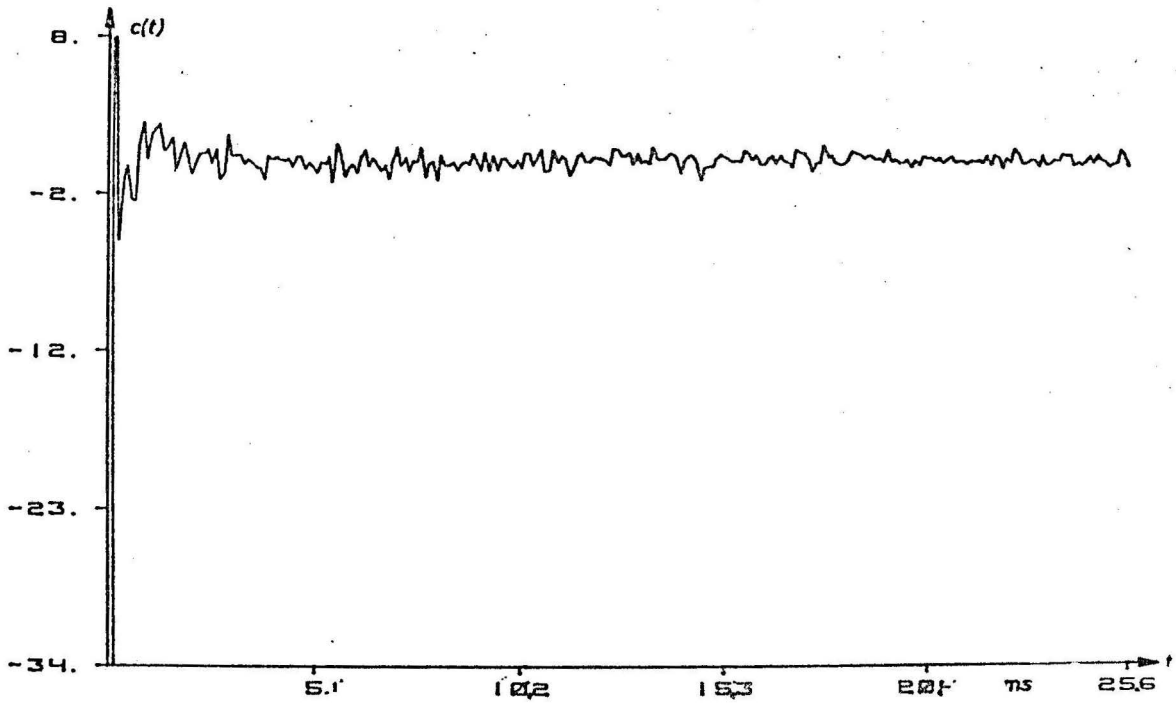


Abb.55b, Cepstrum eines stimmlosen Lautes

des stimmhaften Lautes weist eine auffaellige Spitze bei dem Quefrenzwert auf, der gerade mit der Periodenlaenge des Pulsgenerators der Quelle uebereinstimmt. Im Cepstrum des stimmlosen Lautes nach Abb.55b ist keine ausgepraegte Spitze im Quefrenzbereich der Quelle zu erkennen. Aus dem Vorhandensein oder Nichtvorhandensein der Spitze kann geschlossen werden, ob der Laut stimmhaft oder stimmlos ist. Falls eine solche Spitze vorliegt, kann aus dem zugehoerigen Quefrenzwert die Pitchperiodenlaenge entnommen werden.

Aus den Abtastwerten der Sprache soll der Verlauf der Pitchfrequenz und der Stimmhaft-Stimmlosigkeit ueber der Zeit berechnet werden. Deshalb muss zur Pitchbestimmung das Kurzzeitcepstrum herangezogen werden.

Das Kurzzeitcepstrum wird dadurch berechnet, dass die Sprachzeitfunktion durch ein Zeitfenster, das von Zeitabschnitt zu Zeitabschnitt weitergeschoben wird, bewertet wird. Als Zeitfenster eignet sich z.B. ein sog. Hammingwindow. Es hat den folgenden Zeitverlauf:

$$w(t)=0.54+0.46\cos(\pi\cdot t/T_w) \quad \text{fuer } |t|<T_w \qquad (71)$$
$$w(t)=0 \quad \text{fuer } |t|>T_w$$

Das Zeitfenster muss einerseits kurz sein, um eine hohe Zeitaufloesung zu erhalten, andererseits muss es laenger als die groesste zu ermittelnde Periodenlaenge sein. Vom Verfasser wurde eine Fensterlaenge von 51.2 ms gewaehlt. Aus der mit dem Zeitfenster bewerteten Zeitfunktion wird nach Gl.(68) das Kurzzeitcepstrum berechnet.

Da die Fouriertransformierte des Leistungsspektrums vom Zeitfenster, wie Gl.(72) zeigt, gleich dem Faltungsprodukt

$$\mathcal{F}[|W(\omega)|^2] = \mathcal{F}[W(\omega) \cdot W(-\omega)] = w(t) * w(-t) \quad (72)$$

des Zeitfensters mit sich selbst ist, folgert NOLL /16/ daraus, dass die hochquefrenten Komponenten im Cepstrum, aehnlich wie das Faltungsprodukt der Fensterfunktion mit sich selbst abfaellt. Aus diesem Grunde wird das Cepstrum fuer die nachfolgende Auswertung in dem zu untersuchenden Bereich linear gewichtet. Anfangs- und Endwert der Gewichtkurve wurden experimentell aus den Cepstren von Pulsfolgen unterschiedlicher Pulsfolgefrequenz ermittelt.

Nach NOLL wird aus den gewichteten Kurzzeitcepstren durch Vorgabe einer absoluten Schwelle im Bereich von 1 ms bis 15 ms nach der fuer einen stimmhaften Laut charakteristischen Spitze gesucht.

Am Ende eines stimmhaften Lautes liegt in den meisten Faellen auch ein Abfall der Amplitudenhoehe der Zeitfunktion vor. Da die cepstralen Spitzen in ihrer Groesse mit der Amplitude der Zeitfunktion zusammenhaengen, wuerden diese dann unter die Schwelle fallen und der Laut viel zu frueh als stimmlos erkannt werden. Deshalb wird im Bereich ± 1 ms in der Umgebung der jeweils im vorangegangenen Kurzzeitcepstrum gefundenen Spitze die Schwelle fuer so viele der folgenden Samples herabgesetzt, wie auf stimmhaft entschieden wird. Nach der naechsten Stimmlos-Entscheidung wird die Schwelle wieder heraufgesetzt.

Das Verfahren nach NOLL hat zwei grosse Nachteile:

1. Es benoetigt sehr viel Rechenzeit, da bei der vorgegebenen Fensterlaenge pro Kurzzeitcepstrum zwei Fouriertranformationen fuer jeweils 512 komplexe Werte durchgefuehrt werden muessen.
2. Durch Einsetzen einer absoluten Schwelle werden in vielen Faellen die gesuchten Spitzen nicht gefunden, wenn sie in ihrer Amplitude zu klein sind und es werden dann falsche Spitzen gefunden, wenn mehrere Werte des Cepstrums die Schwelle ueberschreiten.

Nach SANITER /37/ ist die Cepstralanalyse in diesen beiden Punkten verbessert worden.

Die Sprache ist mit einer Abtastfrequenz von 10 kHz abgetastet worden um, unter Beruecksichtigung der in 4.2 erwaehnten Faltungseinfluesse, bei der Analyse noch Frequenzen bis ca 3 kHz verarbeiten zu koennen. Bei der Pitchbestimmung geht es jedoch darum, Frequenzen von hoechstens 200 Hz zu bestimmen. Zu diesem Zweck haette mit einer wesentlich niedrigeren Abtastfrequenz gearbeitet werden koennen und die Fouriertransformation brauchte in dem Fall lediglich auf einen Bruchteil der Werte angewendet zu werden.

Die Tabelle 8 enthaelt eine Aufstellung ueber die gemessenen Rechenzeiten des verwendeten FFT-Programms in Abhaengigkeit von der Anzahl der transformierten komplexen Werte.

Rechenzeit (sek)	komplexe Werte
0.020	16
0.089	32
0.207	64
0.477	128
1.080	256
2.411	512

Tabelle 8, Rechenzeiten des verwendeten FFT-Programms

Daraus ergibt sich, dass die Rechenzeit betraechtlich sinkt, wenn man nicht nur jeden zweiten, sondern jeden vierten oder gar achten Abtastwert verarbeitet. Die Zeitfunktion muss natuerlich vorher durch einen Tiefpass entsprechend niedriger Grenzfrequenz gefiltert werden.

Ein geringfuegiger Nachteil ergibt sich theoretisch daraus, dass bei 10 kHz Abtastfrequenz die Pitchperiode auf 0.1 ms genau bestimmt werden kann, nimmt man dagegen nur jeden achten Abtastwert, kann sie nur noch auf 0.8 ms genau bestimmt werden. Wie sich aus Synthesebeispielen ergeben hat, wirkt sich dieser Nachteil praktisch ueberhaupt nicht auf die Qualitaet der synthetischen Sprache aus.

Der zweite Nachteil der Pitchbestimmung nach NOLL lag in der Verwendung einer absoluten Schwelle fuer die Erkennung der Spitze im Cepstrum. Nach SANITER werden die relativen Lagen

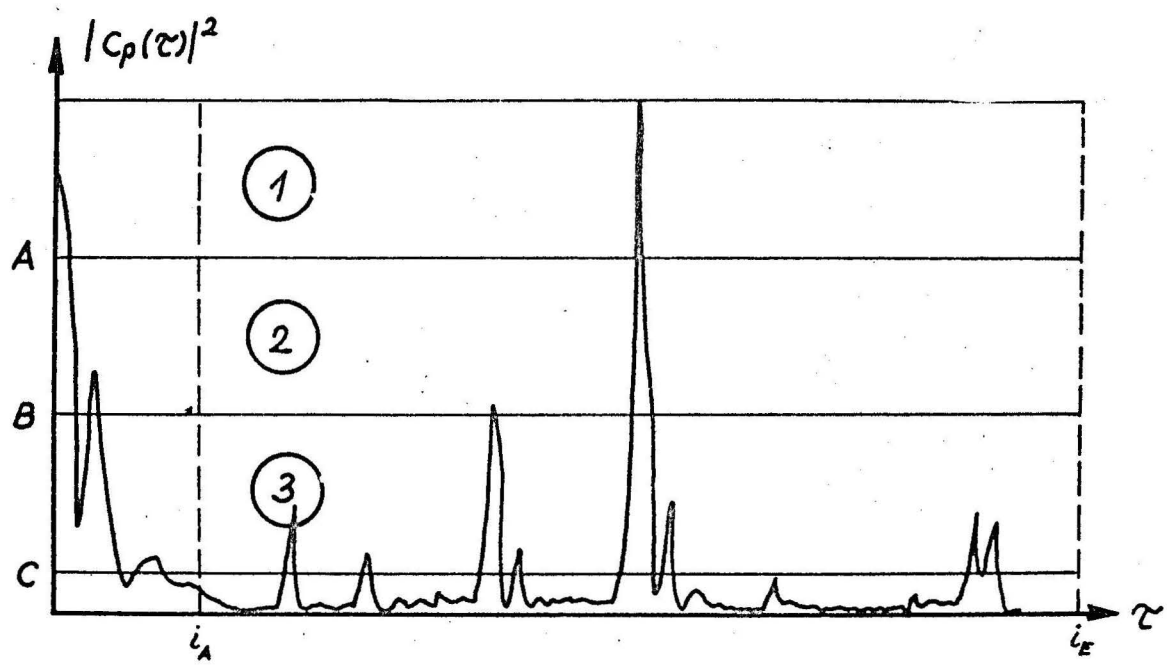


Abb.56, Klasseneinteilung der cepstralen Spitzen

aller im untersuchten Cepstralbereich befindlichen Spitzen zueinander als Entscheidungsmerkmal fuer die Stimmhaftigkeit oder Stimmlosigkeit verwendet. Ausgehend von der groessten Spitze im Kurzzeitspektrum werden die, wie oben bereits erwaehnt, linear gewichteten Amplitudenwerte der Spitzen im Kurzzeitcepstrum in drei Klassen geteilt, so wie es aus der Abb.56 ersichtlich ist. Die unteren Grenzen der Klassen sind

A=7/10 PEAK
B=4/10 PEAK
C=1/10 PEAK

wobei PEAK die Amplitude der groessten Spitze ist.

Aus der Besetzungsanzahl der drei Klassen, dem Mittelwert der groessten Maxima der letzten vier Kurzzeitcepstren und dem Minimum des Mittelwertes wird nach einem auf Erfahrungswerte aufgebauten Algorithmus eine Vorentscheidung fuer die Stimmhaftigkeit oder Stimmlosigkeit des betreffenden Samples getroffen. Die endgueltige Entscheidung wird erst gefaellt, wenn fuer die folgenden Schritte die Stimmhaft-Stimmlosigkeit ermittelt wurde. Es wird eine Glaettung des Parameterverlaufes erreicht, die ein zu haeufiges Wechseln der stimmhaften und stimmlosen Teile der Sprache verhindert, was ohnehin nur auf den Ungenauigkeiten des Entscheidungsalgorithmus beruht.

Die Pitchfrequenz wird, wie bei dem Verfahren von NOLL, aus dem Ort des groessten Maximums im Cepstrum berechnet. Sind zwei etwa gleichgrosse Maxima vorhanden, wird der Wert genommen, der naeher am Frequenzwert des vorangegangenen Samples liegt.

Da die Fouriertransformationen i.a. viel Rechenzeit benoetigen, bringt es eine grosse Zeitersparnis mit sich, wenn man die Zwischenraeume zwischen den Sprachlauten erkennen und bei der Pitchbestimmung ueberspringen kann. Waehrend die stimmlosen Laute durch den Parameter LVU=0 und die stimmhaften Laute durch LVU=1 angezeigt werden, werden die Zwischenraeume mit LVU=2 gekennzeichnet.

Da aus dem Sprachsignal bereits waehrend des Einlesens der Abtastwerte vom Magnetband ein etwaiger Gleichanteil herausgefiltert wird, besteht das Signal in den Pausen aus einem Rauschen mit kleinen Amplituden. Fertigt man eine Verteilungskurve mit logarithmischem Abszissenmassstab fuer die Effektivwerte eines gesprochenen Satzes an, ergibt sich qualitativ eine Darstellung nach Abb.57. Die ausgepraegte Spitze bei kleinen Amplitudenwerten ruehrt von dem Rauschen in den Pausen her. Der Wert MINAMP ist in der Abb.57 der Effektivwert, der den Bereich der Pause von dem des Sprachsignals trennt. Vor jeder Pitchbestimmung wird der Effektivwert des untersuchten Sprachabschnitts berechnet. Ist der Effektivwert kleiner als MINAMP, wird LVU=2 gesetzt und die Pitchbestimmung fuer das betrachtete Sample uebersprungen.

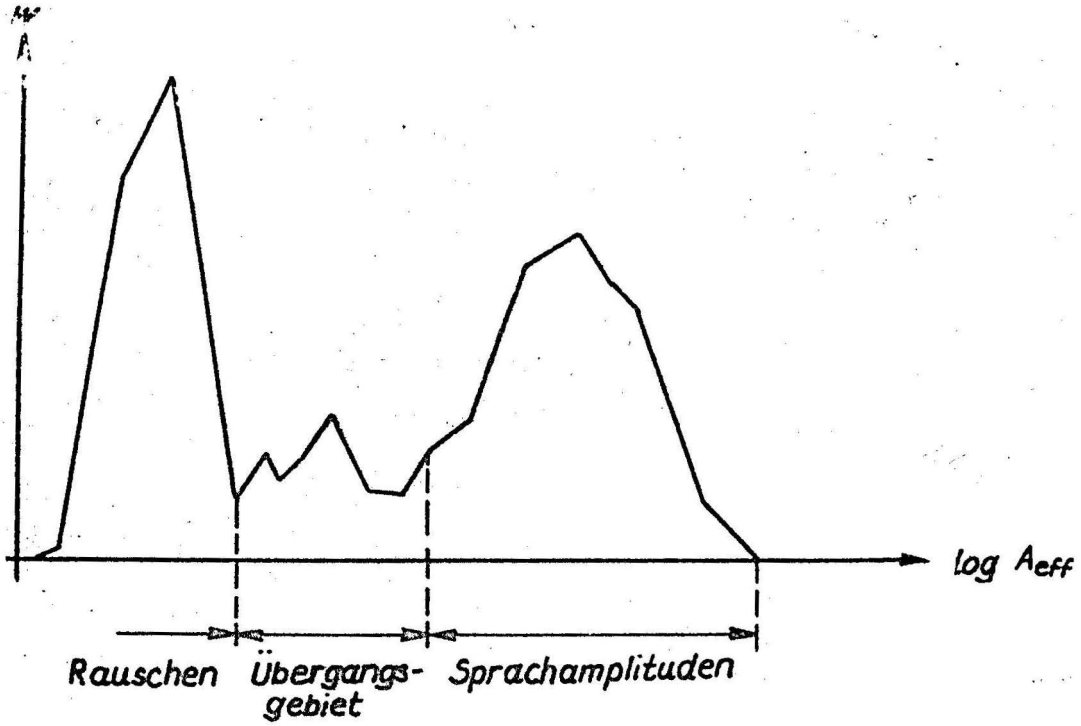


Abb.57, Verteilungskurve fuer die Kurzzeiteffektivwerte von Sprache

Ergebnis

Anhand eines Sprachbeispiels wurde die Pitchbestimmung nach SANITER mit der von NOLL verglichen. Es war keine Qualitaetsminderung der Sprache gegenueber dem Verfahren nach NOLL zu hoeren, selbst wenn nur jeder zweite oder vierte Abtastwert genommen wurde. Lediglich bei der Verwendung jedes achten Abtastwertes war eine geringfuegige Qualitaetsminderung wahrzunehmen, die auf die geringe Anzahl von Werten zurueckzufuehren ist, die in dem Falle zur Auswertung des Cepstrums zur Verfuegung stehen.

6.2.2 Pitchbestimmung aus der Zeitfunktion

In Abb.58a ist ein Ausschnitt aus dem Zeitverlauf eines stimmhaften Lautes wiedergegeben und in Abb.58b der Zeitverlauf eines stimmlosen Lautes. Wie aus den Abbildungen her-

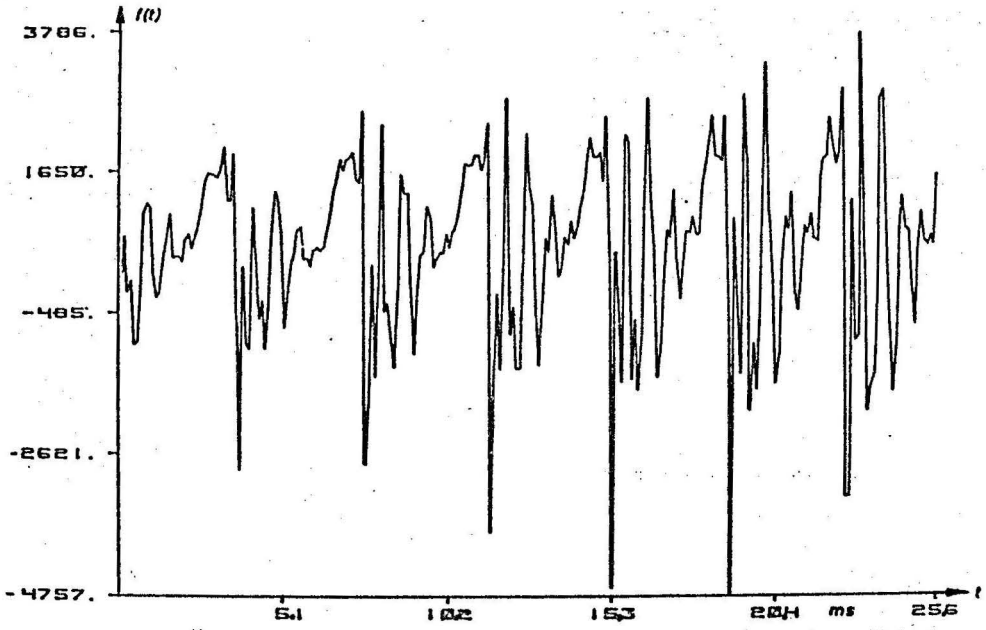


Abb.58a, Zeitfunktion eines stimmhaften Lautes

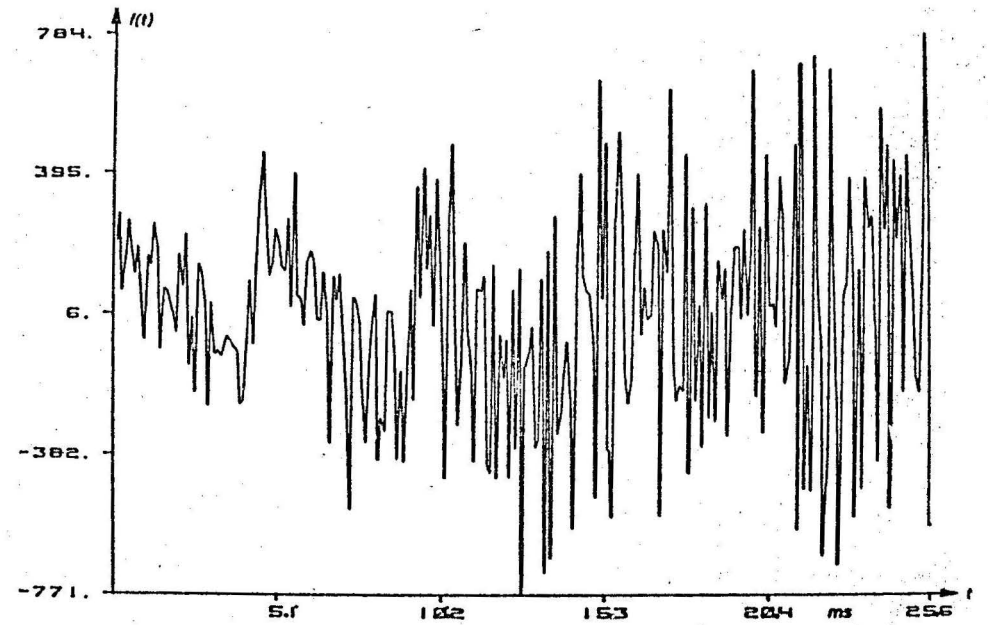


Abb.58b, Zeitfunktion eines stimmlosen Lautes

vorgeht, ist der Zeitverlauf eines stimmhaften Lautes durch eine quasiperiodische Struktur gekennzeichnet. Die Laenge der einzelnen Pitchperioden laesst sich visuell aufgrund auffaelliger Maxima erkennen, die in regelmaessigen Abstaenden auftreten. Bei einer automatischen Pitchfrequenzbe-

stimmung, die auf der genannten Struktur der Sprachzeitfunktion basiert, muessen Kriterien gefunden werden, die die oben erwaehnten auffaelligen Maxima beschreiben. Der Abstand zweier Maxima gibt dann gerade die Laenge der gesuchten Pitchperiode an.

Einen heuristischen Weg, diese sog. signifikanten Peaks zu finden beschreibt REDDY /12/:

1. Aus der Sprachzeitfunktion werden in dem betrachteten Bereich die Maxima und Minima herausgesucht.
2. Von diesen werden diejenigen signifikante Maxima bzw. signifikante Minima genannt,
   a) die positiv (Maxima) bzw. negativ (Minima) sind und mindestens 2.5 ms vom vorhergehenden signifikanten Maximum bzw. Minimum entfernt sind.
   b) die groesser als das 0.9-fache des Maximums in dem betrachteten Intervall sind oder, falls das nicht der Fall ist,
   c) die groesser als die lineare Extrapolation aus den beiden vorhergehenden signifikanten Maxima bzw. Minima sind, oder, falls das auch nicht der Fall ist,
   d) die den groessten Wert im Bereich von 13.5ms vom vorhergehenden signifikanten Maximum bzw. Minimum darstellen.
3. Von den signifikanten Maxima werden die als signifikante Peaks ausgewaehlt, die im Abstand von 3.5 ms ein signifikantes Minimum aufweisen.

Nicht in jedem Fall markieren die nach dieser Vorschrift gefundenen signifikanten Peaks die gewuenschten Stellen der Sprachzeitfunktion. Teilweise werden Markierungen an der falschen Stelle, teilweise zusaetzliche Markierungen gefunden und in manchen Faellen werden Markierungen auch ausgelassen. Es ist deshalb notwendig, nachtraeglich eine Korrektur durchzufuehren.

Die Korrektur basiert auf der Annahme, dass benachbarte Pitchperioden nie mehr als 20% voneinander abweichen. Bezeichnet man die Differenz zwischen dem untersuchten Peak und dem letzten signifikanten Peak als IST-Periode und setzt sie in Beziehung zur letzten berechneten Periode, der sog. SOLL-Periode, so kann man den relativen Fehler

$$RE=(IST-SOLL)/SOLL \quad (73)$$

definieren. Die Korrektur wird mit Hilfe von RE nach einem Algorithmus durchgefuehrt, der in Abb.59 dargestellt ist. Der Korrekturalgorithmus hat die Eigenschaft, dass er auf jeden Fall eine Pitchperiode findet und sogar, falls keine Pitchperiode zu finden ist, durch Einsetzen von Marken eine Pitchperiode erzeugt. Die Ergebnisse dieses Markierungsverfahrens duerfen daher nicht kritiklos hingenommen werden.

KOSSAK /36/ hat festgestellt, dass die Bestimmung der Pitchmarken nach REDDY zu einer Anhaeufung kleiner Periodenlaengen, insbesondere von 2.5 ms-Perioden fuehrt, die dem Mindestabstand zweier benachbarter Marken entsprechen. Die

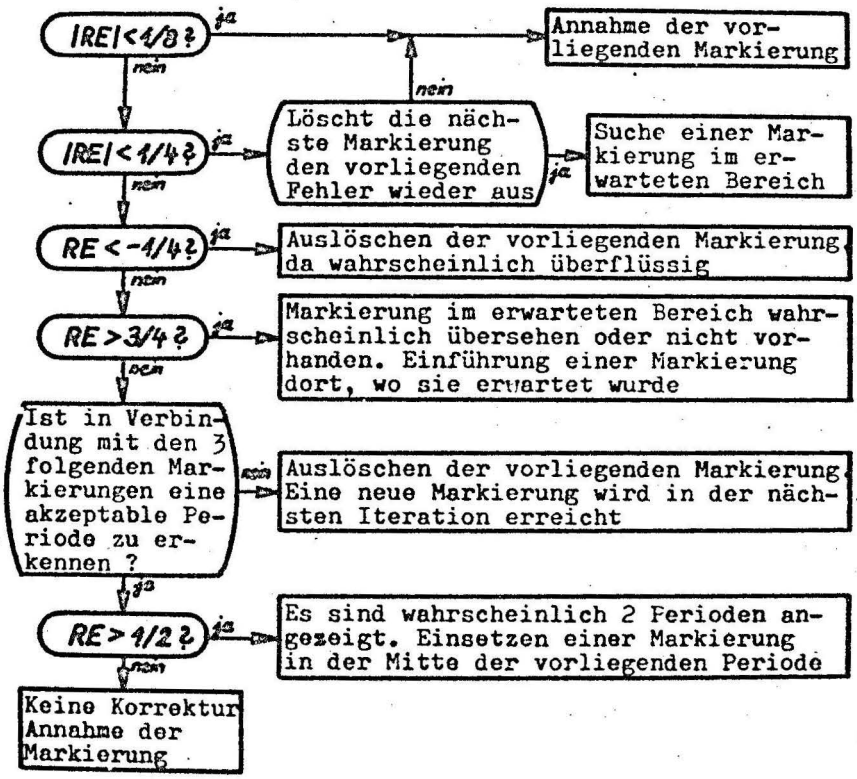


Abb.59, Korrekturalgorithmus nach REDDY

Gruende dafuer sind die folgenden:

1. Es koennen in dem betrachteten Bereich mehrere Maxima bzw. Minima auftreten, die betragsmaessig groesser als das 0.9-fache des Spitzenwertes sind. Sie werden nach der Signifikanzbedingung 2b) zu signifikanten Maxima bzw. Minima erklaert. Durch eine Erhoehung des Faktors 0.9 auf beispielsweise 0.94 liesse sich das Ergebnis in diesem Fall verbessern. Die Erhoehung des Faktors auf 0.94 fuehrt aber andererseits dazu, dass an anderen Stellen die gesuchten Peaks nicht mehr erkannt werden koennen. Deshalb wurde der Faktor 0.9 im vorliegenden Programm beibehalten.

2. Nach Punkt 2d) der Signifikanzbedingung wird selbst ein alleine vorhandenes lokales Maximum zum signifikanten Peak erklaert, wenn es weniger als 13.5 ms von der vorhergehenden Marke entfernt ist. Ist dieses Maximum recht klein, wird beim naechsten Schritt entsprechend der Signifikanzbestimmung 2c) durch Extrapolation mit Sicherheit eine neue 2.5ms-Periode konstruiert. Eine geringfuegige Verbesserung fuer diesen Fall konnte dadurch erreicht werden, dass die Suche nach dem groes-

sten lokalen Maximum bzw. Minimum an dieser Stelle nur eingeleitet wird, wenn mindestens 4 Werte zur Auswahl zur Verfuegung stehen.

3. Es werden zwei signifikante Maxima im Abstand von 3 ms durch ein signifikantes Minimum zu zwei Markierungen gemacht, wenn das Minimum gerade in der Mitte zwischen den Maxima liegt.

Ein weiterer kritischer Punkt in der Bestimmung der Periodenlaenge nach REDDY ist die Berechnung der SOLL-Periode, denn sie hat eine zentrale Bedeutung fuer die Handhabung des Korrekturalgorithmus. Die SOLL-Periode ist im vorliegenden Programm gleich der vorhergehenden Periode. Sie koennte aber auch durch Extrapolation oder Mittelwertbildung aus den vergangenen Periodenlaengen ermittelt werden. Die Gefahr ist in jedem Falle, dass die Ermittlung einer falschen Periodenlaenge eine falsche Markierung zur Folge hat, und dass dieser Fehler sich immer weiter fortpflanzt.

Stimmhaft-Stimmlos-Entscheidung

Eine Moeglichkeit die Stimmhaft-Stimmlos-Entscheidung durchzufuehren ergibt sich aus der Signifikanzbestimmung unter dem Gesichtspunkt, dass bei stimmlosen Lauten entweder keine Markierungen gefunden werden koennen oder die gefundenen Pitchperioden in ihrer Laenge regellos verteilt sind. Um die Regellosigkeit festzustellen, bieten sich zwei Moeglichkeiten an:

1. Die ermittelten Laengen der Pitchperioden werden in eine endliche Anzahl von Klassen eingeteilt. Sind nur wenige Klassen, insbesondere eine einzige Klasse besetzt, ist der Laut mit grosser Wahrscheinlichkeit stimmhaft. Damit ist die Anzahl der besetzten Klassen ein Richtwert fuer die Stimmhaft-Stimmlos-Entscheidung.

2. Es wird die Streuung der ermittelten Pitchperiodenlaengen berechnet. Bei grosser Streuung ist der Laut stimmlos, bei kleiner Streuung stimmhaft.

Es zeigt sich, dass die besten Ergebnisse durch Kombination dieser beiden Moeglichkeiten in Form des sog. Streuproduktes erzielt werden koennen. Das Streuprodukt besteht aus dem Produkt der Gesamtstreuung der Markierungen im Arbeitsbereich mit der Anzahl der belegten Klassen. Als Arbeitsbereich wurde eine Zeitspanne von 80 ms gewaehlt.

Der Schwellenwert, der bei der Ueber- oder Unterschreitung durch das Streuprodukt zu einer Stimmlos- oder Stimmhaftentscheidung fuehrt, wird bei einem Uebergang von stimmlos nach stimmhaft hoch und beim Uebergang von stimmhaft nach stimmlos niedrig angesetzt. Durch diese Massnahme soll der Einfluss des Korrekturprogramms kompensiert werden, das, wie bereits oben erwaehnt wurde, auch in stimmlosen Bereichen Markierungen erzeugen kann.

Fuer die Zwischenraeume zwischen zwei Lauten wird der Parameter LVU, der die Stimmhaftigkeit mit LVU=1 und die Stimmlosigkeit mit LVU=0 darstellt, LVU=2 gesetzt. Das ist die sog. Pausenentscheidung. Das Vorhandensein einer Pause wird dadurch erkannt, dass in dem zu charakterisierenden Bereich von 10 ms, also bei 100 Abtastwerten, eine vorgegebene Schwelle hoechstens dreimal ueberschritten wird.

Ergebnis

Das Verfahren ist wesentlich schneller als die Pitchbestimmung ueber das Cepstrum, da keine Fouriertransformation benoetigt wird. Das Verfahren weist aber andererseits beim akustischen Vergleich mit den Verfahren nach SANITER /37/ und TULGAN /39/ deutlich hoerbare Maengel auf. Es hat sich vor allem gezeigt, dass das Verfahren sehr empfindlich gegen Verschiebungen der Nullage der Sprachzeitfunktion ist.

Zum gegenwaertigen Zeitpunkt wird noch untersucht, inwieweit sich die Qualitaet der Pitchanalyse verbessern laesst, wenn man die Sprache durch einen Spectrum-flattener vorverzerrt.

6.3 Formantbestimmung im Frequenzbereich

6.3.1 Berechnung der Kurzzeituebertragungsfunktion

Nach Gl.(12) in Kap 2.2 wird die Uebertragungsfunktion des Vokaltraktes, insbesondere bei der Erzeugung stimmhafter Laute durch Formanten charakterisiert. Will man die Formanten aus dem Frequenzbereich berechnen, empfiehlt es sich, zur Analyse den Verlauf des Betrages der Uebertragungsfunktion und nicht einfach das Spektrum der Sprache heranzuziehen.
Es gibt im wesentlichen drei Moeglichkeiten, den Betrag der Uebertragungsfunktion zu berechnen:

1. Das Kurzzeitspektrum der Zeitfunktion wird dadurch berechnet, dass die Zeitfunktion mit einem Zeitfenster multipliziert und das Produkt mit dem FFT-Algorithmus in den Frequenzbereich transformiert wird.
   Die Zeitfunktion ist eine reelle Funktion. Im Frequenzbereich erhaelt man i.a. eine komplexe Funktion. Durch Betragsbildung kann man aus der komplexen Funktion den Betrag des Spektrums berechnen.
   Das Kurzzeitspektrum besteht ebenso, wie Gl.(19) in Kap 2.4 es beschreibt, aus den Anteilen der Quelle, dem Formantfilter und einem Korrekturglied, das die Abstrahlung beruecksichtigt. Die Abstrahlung und der Anteil des Formantfilters erscheinen als Huellkurve des Kurzzeitspektrums, waehrend die Quelle durch einen periodischen Ripple gekennzeichnet ist, der aus allen Vielfachen der Pitchfrequenz besteht. Der Einfluss der Abstrahlung kann aus dem logarithmierten Spektrum durch Subtraktion einer Korrekturkurve entfernt werden. Dann besteht das Restspektrum nur noch aus den Anteilen der Quelle und denen des Formantfilters.
   Die Frequenzaufloesung im Spektrum wird durch die Breite des verwendeten Zeitfensters bestimmt. Waehlt man ein schmales Zeitfenster, z.B. T=5 ms, werden Frequenzen unter f=1/T=200 Hz nicht mehr aufgeloest. Das bedeutet, dass auch der Einfluss der Quelle nicht mehr aufgeloest wird. Bei Verwendung eines schmalen Zeitfensters entspricht, unter Beruecksichtigung der Abstrahlung, das Kurzzeitspektrum dem gesuchten Betrag der Kurzzeituebertragungsfunktion des Formantfilters.
   Ein Nachteil dieser Methode ist, dass die Kurzzeituebertragungsfunktion nur durch wenige Werte, d.h. ungenau dargestellt wird.

2. Die zweite Moeglichkeit schliesst sich an den Gedankengang der ersten an. Das gleiche Resultat wie unter 1. kann naemlich dadurch erreicht werden, dass man zur Ermittlung der Kurzzeituebertragungsfunktion des Vokal-

traktes eine Filterbank heranzieht. Die Bandbreite der einzelnen Bandpaesse muss dann groesser als die Pitchfrequenz sein, damit diese im Spektrum nicht mehr aufgeloest werden kann. Die Amplitudenwerte an den Ausgaengen der Bandpaesse ergeben gleichgerichtet und durch Tiefpaesse geglaettet die gesuchten Spektralwerte.

3. Fuer genauere Untersuchungen der Kurzzeituebertragungsfunktion wird eine hoehere Frequenzaufloesung benoetigt, als das in 1. und 2. der Fall ist. In diesem Fall muss ein breiteres Zeitfenster verwendet werden. Der Verfasser benuetzte in den meisten Faellen ein Hamming-Window mit einer Gesamtlaenge von 51.2 ms und erhielt damit eine Frequenzaufloesung von 19.5 Hz.
Der Einfluss der Quelle kann durch homomorphe Filterung entsprechend Kap 6.1 aus dem Spektrum herausgefiltert werden. Es wird zunaechst das Cepstrum berechnet und dort der hochquefrente Anteil, der das Cepstrum der Quelle darstellt, herausgeschnitten. Durch Ruecktransfomation des Restcepstrums gelangt man wieder in den Spektralbereich. Dort muss die Korrekturkurve, die den Einfluss der Abstrahlung beruecksichtigt, subtrahiert werden.
Damit liegt der logarithmierte Betragsverlauf der Kurzzeituebertragungsfunktion des Vokaltraktes vor, der eventuell durch Exponentiation auf den linearen Massstab zurueckgefuehrt werden kann.

6.3.2 Formantbestimmung aus spektralen Momenten

Da der Syntheseteil des Formantvocoders alle 10 ms einen vollen Parametersatz zur Steuerung benoetigt, muss auch fuer jedes 10 ms-Intervall der Sprache eine Formantbestimmung durchgefuehrt werden. Das bedeutet, dass das Zeitfenster, mit dem die Zeitfunktion zur Berechnung des Kurzzeitspektrums bewertet werden muss, in 10 ms-Intervallen weitergeschoben wird. Das Zeitfenster hat fuer den vorliegenden Fall einen $\cos^2$ -Verlauf und eine Laenge von insgesamt 25.6 ms. Das Spektrum wird durch die diskrete Fouriertransformation mit dem FFT-Algorithmus aus den mit dem Zeitfenster gewichteten Abtastwerten berechnet. Das Spektrum, das aufgrund der Abtastfrequenz von 10 kHz bis zu einer Frequenz von 5 kHz berechnet wird, wird durch 128 diskrete Spektralwerte, die sog. Kanaele dargestellt. Der Abstand zweier benachbarter Kanaele betraegt 39 Hz.

Da fuer die in der Sprache auftretenden Formanten die Bedingung

$$\omega_p \gg \sigma_p$$

erfuellt ist, treten, wie in Abb.60 dargestellt ist, im

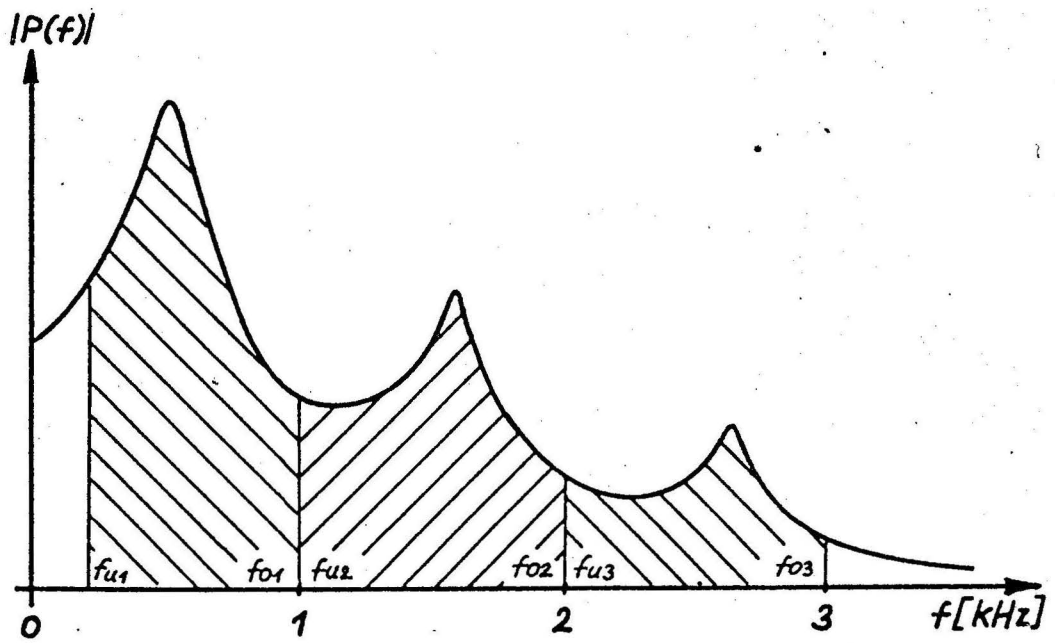


Abb.60, Bereichsgrenzen fuer die Momentbildung

Spektrum an der Stelle der Polfrequenzen Maxima auf. Die Frequenzlagen der Maxima lassen sich durch einfache Moment-

bildung nach Gl.(74) aus dem Bereich in der unmittelbaren Umgebung des betreffenden Maximums recht gut berechnen.

$$f_i = \frac{\sum_{j=u_i}^{o_i} f_j |P(f_j)|}{\sum_{j=u_i}^{o_i} |P(f_j)|} \qquad (74)$$

In der Gl.(74) ist u_i die Kanalnummer fuer die untere Bereichsgrenze und o_i die Kanalnummer fuer die obere Bereichsgrenze. Die f_j sind die zu den Kanaelen gehoerenden Frequenzwerte und $P(f_j)$ die Spektralwerte des Drucks im Schallfeld des Sprechers.

Das Ergebnis einer Formantbestimmung nach Gl.(74) ist um so genauer, je naeher die untere und obere Bereichsgrenze an dem Maximum liegen und je ausgepraegter das Maximum ist.

Um die Berechnung besonders einfach zu gestalten, wurden fuer die gesamte Rechnung konstante Grenzen fuer die einzelnen Formantbereiche angenommen. Die Lagen der Formantbereiche sind in Abb.60 eingezeichnet. Die Grenzen sind:

f_{u1} = 230 Hz f_{o1} =1000 Hz fuer den ersten Formanten
f_{u2} =1000 Hz f_{o2} =2000 Hz fuer den zweiten Formanten
f_{u3} =2000 Hz f_{o3} =3000 Hz fuer den dritten Formanten

Die Maxima im Spektrum werden als solche noch dadurch besonders hervorgehoben, dass zur Momentberechnung die Amplitudenquadrate der Spektralwerte herangezogen werden. Die Berechnung der Formantfrequenzwerte erfolgt somit nach Gl.(75)

$$f_i = \frac{\sum_{j=u_i}^{o_i} f_j |P(f_j)|^2}{\sum_{j=u_i}^{o_i} |P(f_j)|^2} \qquad (75)$$

Die Rechenzeit fuer 1 sek Sprache betraegt auf der CAE 90-40 180 sek.

6.3.3 Momentverfahren nach NAKATSIN und SUZUKI /39/

Bei dem Momentverfahren nach NAKATSIN und SUZUKI wird zur Berechnung des Kurzzeitspektrums ein $\cos^2$-Zeitfenster der Laenge von 12.8 ms verwendet. Das Spektrum bis zu 5 kHz wird in dem Fall durch 64 Kanaele dargestellt, die voneinander einen Abstand von ca 79 Hz haben. Durch die Wahl eines derart schmalen Zeitfensters ist der Einfluss der Quelle auf das Spektrum weitgehend eliminiert worden. Das Spektrum laesst sich daher nach Gl.(76) durch die Spektren $H_i(f)$ der ersten drei Formanten und eine Korrekturkurve KR(f) darstellen.

$$P(f) = \prod_{i=1}^{3} H_i(f) \cdot KR(f) \qquad (76)$$

Durch Logarithmierung der Gl.(76) ergibt sich Gl.(77):

$$\log P(f) = \sum_{i=1}^{3} \log H_i(f) + \log KR(f) \qquad (77)$$

Sind der Verlauf der Korrekturkurve KR(f) und die Lage des ersten und dritten Formanten bekannt, so kann man den Spek-

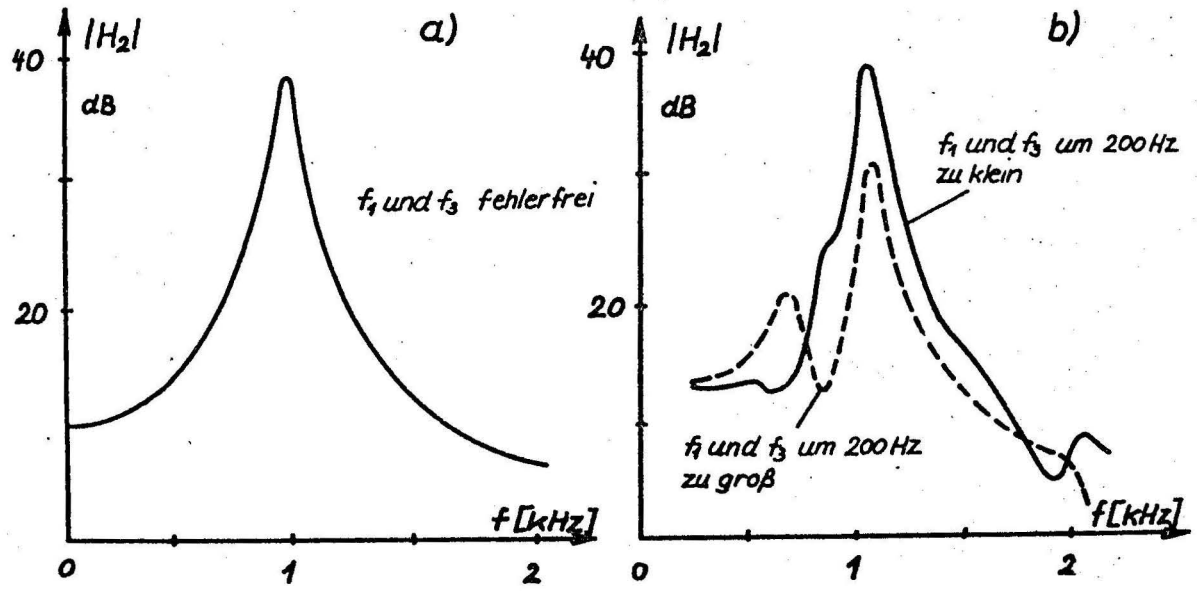


Abb.61, Berechnung eines Formanten aus Gl.(78)
a) bei exakten f1- und f2-Werten
b) bei ungenauen f1- und f2-Werten

tralverlauf des zweiten Formanten aus Gl.(78) berechnen:

$$\log H_2(f) = \log P(f) - \left[\sum_{\substack{i=1 \\ i\neq 2}}^{3} \log H_i(f) + \log KR(f)\right] \qquad (78)$$

Da in den meisten Faellen die Frequenzlagen des ersten und dritten Formanten nicht genau bekannt sind, wird aus Gl.(78) nicht ein Verlauf nach Abb.61a, sondern eher ein Verlauf nach Abb.61b bei der Differenzbildung herauskommen. Bildet man das spektrale Moment erster Ordnung in einer gewissen Umgebung des Maximums entsprechend Gl.(74), so kann man auch bei Spektralverlaeufen nach Abb.61b noch sehr gut den Frequenzwert des zweiten Formanten berechnen.

Der neu berechnete Wert von f_2 kann jetzt zusammen mit f_3 zur Berechnung von f_1 nach Gl.(79) und Gl.(74) bzw. zusammen mit f_1 zur Berechnung von f_3 nach Gl.(80) und Gl.(74) verwendet werden.

$$\log H_1(f) = \log P(f) - \Big[\sum_{\substack{i=1 \\ i\neq 1}}^{3} \log H_i(f) + \log KR(f)\Big] \qquad (79)$$

$$\log H_3(f) = \log P(f) - \Big[\sum_{\substack{i=1 \\ i\neq 3}}^{3} \log H_i(f) + \log KR(f)\Big] \qquad (80)$$

Die ersten drei Formantfrequenzen werden auf diese Weise zyklisch und zwar in der Reihenfolge $f_2 \rightarrow f_3 \rightarrow f_1$ berechnet. Der Zyklus wird dann unterbrochen, wenn die Differenz zwischen den neuesten Frequenzwerten und den im vorangegangenen Zyklus berechneten fuer alle drei Formanten eine vorgegebene Schwelle unterschreitet.

Zu Beginn der zyklischen Berechnung muessen Anfangswerte fuer die ersten drei Formanten vorgegeben werden. Die Anfangswerte sind dabei jeweils die aus dem vorhergehenden Kurzzeitspektrum berechneten Frequenzwerte. Die Anfangswerte, die zur Berechnung des allerersten Samples verwendet werden, sind die Frequenzwerte 500 Hz, 1500 Hz, 2500 Hz.

Die Grenzen des Gebietes, innerhalb dessen die spektralen Momente berechnet werden, sind variabel. Der untere Frequenzwert f_u und der obere Frequenzwert f_o berechnen sich nach Gl.(81) zu:

$$f_u = f_j(1-\alpha) - \beta \qquad \alpha = 0.15 \qquad (81)$$

$$f_o = f_j(1+\alpha) + \beta \qquad \beta = 200\ Hz$$

Der Index j gibt dabei die Nummer des zu berechnenden Formanten an und f_j ist der jeweilige Anfangswert fuer die Berechnung.

Die Rechenzeit betraegt fuer 1 sek Sprache auf der CAE 90-40 300 sek.

6.3.4 Formantbestimmung nach SCHAFER und RABINER (/40/)

Nach SCHAFER und RABINER wird zur Berechnung des Kurzzeitspektrums ein Hammingwindow nach Gl.(71) mit einer Gesamtlaenge von 51.2 ms verwendet. Die Frequenzaufloesung im Spektrum betraegt damit 19.5 Hz. Die Anteile des Spektrums, die von der Quelle herruehren, werden durch homomorphe Filterung mit Hilfe des Cepstrums aus dem Spektrum herausgefiltert. Anschliessend wird vom Spektrum noch eine Korrekturkurve subtrahiert, die den Einfluss der Abstrahlung und den der Glottis beruecksichtigt. Das Kurzzeitspektrum enthaelt damit nur noch den Anteil, der von der Uebertragungsfunktion des Vokaltraktes herruehrt.

Aus den Maxima des Kurzzeitspektrums, d.h. aus deren Frequenzlage, und Amplitudenwerten werden die ersten drei Formanten nach einem vorgegebenen Algorithmus ermittelt. Es wird dabei kein Bezug auf vorher bestimmte Formantwerte genommen, so dass die Formantwerte fuer alle Kurzzeitspektren unabhaengig voneinander bestimmt werden. Dadurch wird verhindert, dass Fehler in der Formantbestimmung aufgrund frueherer Fehlentscheidungen auftreten und sich weiter fortpflanzen koennen.

Die Formantbestimmung nach SCHAFER und RABINER beruecksichtigt im Gegensatz zur Formantbestimmung nach 6.3.1, dass sich die Frequenzbereiche fuer die einzelnen Formanten stark ueberlappen koennen. Aus umfangreichen Untersuchungen fanden SCHAFER und RABINER die folgenden Bereichsgrenzen fuer maennliche Sprecher heraus:

1. Formant: f1MN = 200 Hz bis f1MX = 900 Hz
2. Formant: f2MN = 550 Hz bis f2MX = 2700 Hz
3. Formant: f3MN = 1100 Hz bis f3MX = 2950 Hz

Wie dem zu entnehmen ist, koennen sich der Bereich des ersten und des zweiten Formanten von 550 bis 900 Hz und der Bereich des zweiten und dritten Formanten sich von 1100 bis 2700 Hz ueberlappen.

Ein weiterer Punkt der von SCHAFER und RABINER beruecksichtigt wird, ist das Amplitudenverhaeltnis der spektralen Spitzen. Hierauf wird besonders bei der Unterscheidung zwischen dem ersten und dem zweiten Formanten grosser Wert gelegt. Die Amplitudendifferenz

$$\Delta_{12} = \log|H(e^{j2\pi f_2 T})| - \log|H(e^{j2\pi f_1 T})| \quad (82)$$

ist weitgehend unabhaengig von hoeheren Formanten und ist in Abb.62 (/40/ S.641) als Funktion der Frequenz des zweiten Formanten dargestellt.

Bestimmung des ersten Formanten

Zunaechst wird die groesste Spitze im Bereich von 0 Hz bis f1MX gesucht und die zugehoerige Amplitude als f0AMP abge-

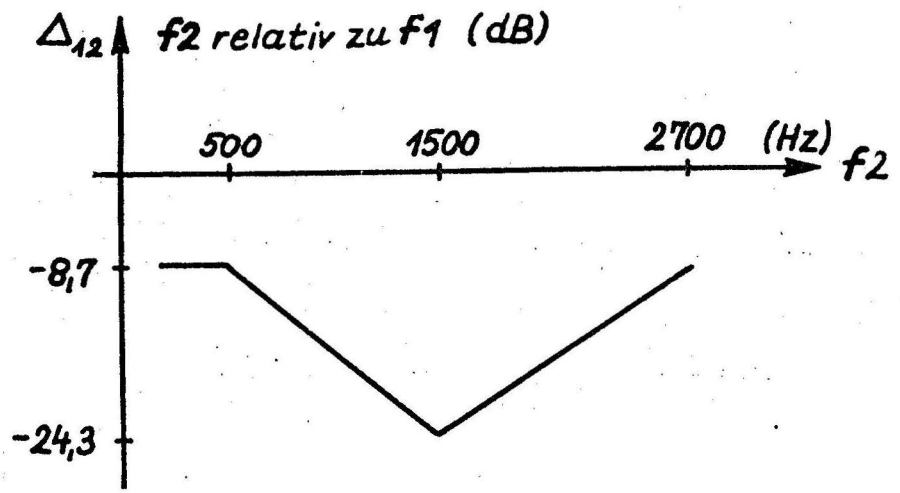


Abb.62, Verlauf von Δ_{12} nach Gl.(82)

speichert. Wenn der Frequenzwert im Bereich des ersten Formanten liegt, wird er sofort als f1, d.h. als erster Formant akzeptiert. Liegt die groesste Spitze unterhalb von f1MN, wird im Bereich von f1 das groesste Maximum gesucht, das aber hoechstens 8.7 dB kleiner als f0AMP sein darf. Der Amplitudenwert dieses Maximums wird f1AMP genannt.

Sollte sich immer noch keine Spitze finden, die den gestellten Anforderungen genuegt, wird der Spektralbereich von 0 bis 900 Hz mit Hilfe der Chirp-z-Transformation vergroessert und bezueglich der Polwerte entdaempft dargestellt. Der erste Formant stellt dann die groesste Spitze im f1-Bereich dar.

Sollte sich auch jetzt noch kein Kandidat fuer den ersten Formanten gefunden haben, wird als erster Formant der Frequenzwert f1MN und als zugehoeriger Amplitudenwert

$$f1AMP = f0AMP - 8.7\ dB$$

angenommen.

Bestimmung des zweiten Formanten

Zuerst muss der Frequenzbereich ermittelt werden, in dem der zweite Formant gesucht werden soll. Wenn f1 kleiner als f2MN ist, wird im Bereich f2MN bis f2MX gesucht. Ist f1 groesser als f2MN, kann es moeglich sein, dass der erste Formant in Wirklichkeit der zweite Formant ist. Deshalb muss der Be-

reich von f1MN bis f2MX nach dem zweiten Formanten durchsucht werden.

Zur Ermittlung von f2 wird die Kurve nach Abb.62 benutzt. Es wird die Spitze als zweiter Formant akzeptiert, bei der die Differenz zwischen f1AMP und dem Amplitudenwert der betreffenden Spitze die Schwellenkurve nach Abb.62 am hoechsten ueberragt. f2AMP ist dann der zugehoerige Amplitudenwert. Liegt die Differenz dagegen unter der Schwellenkurve, scheidet der betreffende Wert als Kandidat aus.

Kann der zweite Formant auf diesem Wege nicht ermittelt werden, liegt die Vermutung nahe, dass der erste und der zweite Formant so dicht beieinanderliegen, dass sie im Spektrum nicht mehr als getrennte Spitzen aufgeloest werden koennen. Im Bereich ±450 Hz um f1 wird dann das Spektrum vergroessert und entdaempft mit der Chirp-z-Transformation dargestellt. Werden dabei zwei benachbarte Spitzen festgestellt, ist die mit dem niedrigeren Frequenzwert der erste Formant und die mit dem hoeheren Frequenzwert der zweite Formant. Wenn aber auch durch die Chirp-z-Transformation keine Aufloesung der beiden dicht benachbarten Spitzen erreicht werden kann, wird der zweite Formant im Abstand von 200 Hz vom ersten Formanten angenommen.

Bestimmung des dritten Formanten

Wenn der zweite Formant in seiner Frequenz groesser als f3MN ist, besteht die Moeglichkeit, dass der zweite und dritte Formant vertauscht wurden und es wird deshalb der Bereich von f2MN bis f3MX durchsucht. Anderenfalls wird der dritte Formant lediglich im Frequenzbereich von f3MN bis f3MX gesucht. Es wird fuer alle Maxima die Differenz zu f2AMP, dem Amplitudenwert des zweiten Formanten, gebildet. Es wird der Wert als dritter Formant genommen, bei dem die Differenz einen Schwellenwert von 17.4 dB am weitesten unterschreitet. Konnte keine derartige Spitze gefunden werden, muss man annehmen, dass der zweite und dritte Formant so dicht beieinander liegen, dass sie als ein Maximum erscheinen. Der Bereich ±450 Hz um den zweiten Formanten wird dann mit Hilfe der Chirp-z-Transformation wiederum vergroessert und entdaempft dargestellt. Wenn dabei zwei Spitzen festgestellt werden koennen, stellt der niedrigere Frequenzwert den zweiten Formanten und der hoehere Frequenzwert den dritten Formanten dar.

Fuer den Fall, dass auch bei Anwendung der Chirp-z-Transformation keine zwei Spitzen aufgeloest werden konnten, wird der Frequenzwert des dritten Formanten als 200 Hz ueber dem des zweiten Formanten angenommen. In jedem Fall wird zum Schluss geprueft, ob f1 < f2 < f3 ist. Sollte das an einer Stelle nicht der Fall sein, werden die entsprechenden Werte miteinander vertauscht.
Die Rechenzeit betraegt auf der CAE 90-40 fuer 1 sek Sprache ca 1200 sek.

6.3.5 Halbautomatische Formantbestimmung mit dem Display

Mit diesem Verfahren (/41/,/42/) koennen sowohl Formanten als auch Antiformanten nach einem 'Analysis by Synthesis'-Prinzip aus der Uebertragungsfunktion des Vokaltraktes bestimmt werden.

Die Uebertragungsfunktion des Vokaltraktes wird, wie bereits in 6.3.3 beschrieben wurde, durch homomorphe Filterung aus dem Kurzzeitspektrum der Sprachzeitfunktion gewonnen. Als Zeitfenster fuer die Berechnung des Kurzzeitspektrums wird ein Hammingwindow nach Gl.(71) mit einer Laenge von 51.2 ms verwendet. Die Uebertragungsfunktion des Vokaltraktes wird daher durch 256 diskrete Werte dargestellt. Die Frequenzaufloesung betraegt 19.5 Hz.
Die Uebertragungsfunktion des Vokaltraktes wird in logarithmischem Amplitudenmassstab auf dem Display graphisch dargestellt. Der Operateur entscheidet, in welchem Bereich ein Formant oder ein Antiformant gesucht werden soll und markiert den betreffenden Bereich mit der Lichtpistole auf dem Schirm. Der Rechner berechnet den in dem markierten Bereich befindlichen Formanten oder Antiformanten nach dem unten beschriebenen Verfahren nach Frequenz und Bandbreite. Er subtrahiert anschliessend die logarithmierte Uebertragungsfunktion des gefundenen Formanten oder Antiformanten von der logarithmierten Uebertragungsfunktion des Vokaltraktes. Der Operateur kann dann aus der Restkurve den naechsten Forman-

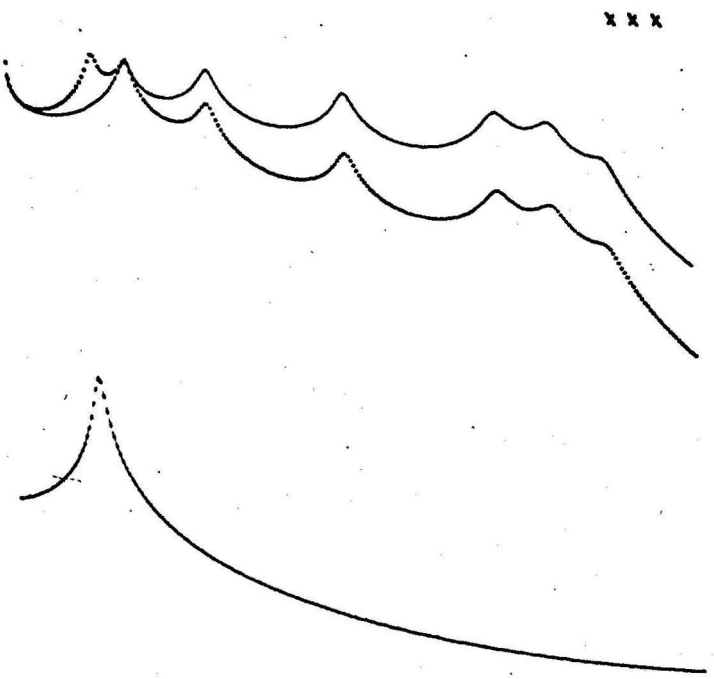


Abb.63, Ausschnitt aus der Formantanalyse
Bestimmung des ersten Formanten

ten oder Antiformanten heraussuchen und so in mehreren Schritten die vorgegebene Uebertragungsfunktion sehr genau analysieren. Die Abb.63 zeigt einen Zeitpunkt der Formantanalyse. Der gerade bestimmte erste Formant ist in Abb.63 unten zu sehen. Die oberste Kurve stellt den Frequenzgang dar, der analysiert werden soll und die mittlere Kurve zeigt den gleichen Frequenzgang, von dem der gefundene Pol subtrahiert worden ist.

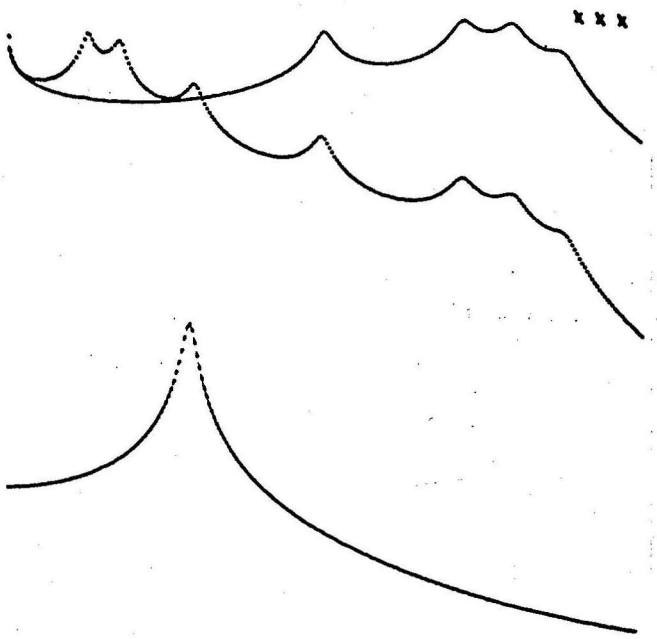


Abb.64, Ausschnitt aus der Formantanalyse
Bestimmung des dritten Formanten

Abb.64 zeigt den Ausschnitt der Analyse, bei dem gerade der dritte Formant bestimmt wird. Man sieht unten den dritten Formanten, in der Mitte den zu analysierenden Frequenzgang und oben den Restverlauf, von dem bereits die ersten drei Formanten subtrahiert worden sind.

Arbeitsweise des Programms

Eine Uebertragungsfunktion wird allgemein nach Gl.(5) beschrieben. Mit der Einfuehrung eines Formanten nach Gl.(6) und eines Antiformanten nach Gl.(7) ergibt sich die Uebertragungsfunktion, die hier mit T(s) bezeichnet werden soll, zu:

$$T(s) = \prod_{i=1}^{n} H_{pi} \cdot \prod_{j=1}^{m} H_{zj} \qquad (83)$$

Die logarithmische Darstellung, die nach Gl.(84) eingefuehrt

$$\left.\begin{aligned} T_{dB}(\omega) &= 20\sigma \cdot lg\,|T(s)| \\ G_{dB}(\omega) &= 20 \cdot lg\,|H_{pi}(s)| \\ H_{dB}(\omega) &= 20\ lg\,|H_{zi}(s)| \end{aligned}\right\} \quad (84)$$

wird, fuehrt zur Gl.(85).

$$T_{dB}(\omega) = \sum_{i=1}^{n} G_{dB_i}(\omega) + \sum_{j=1}^{m} H_{dB_j}(\omega) \quad (85)$$

Diese Gleichung, in der Formanten und Antiformanten additiv verknuepft sind, ist der Ausgangspunkt fuer die vorliegende Formantanalyse.

Frequenzgang eines Formanten und eines Antiformanten

Wenn in diesem Kapitel von 'Frequenzgang' gesprochen wird, soll damit automatisch immer die logarithmische Darstellung entsprechend Gl.(84) und Gl.(85) gemeint sein.

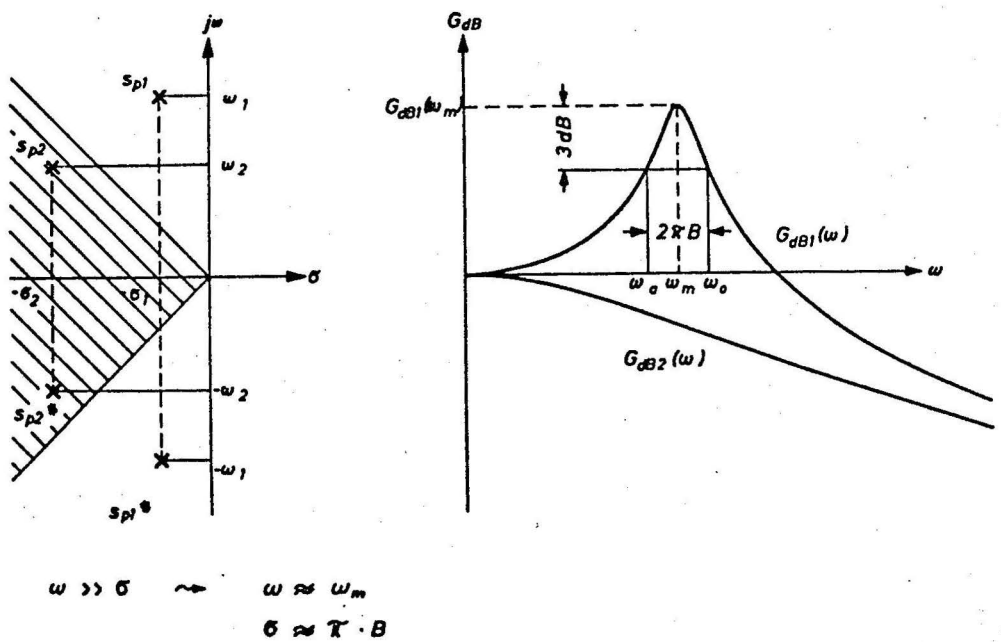


Abb.65, Frequenzgaenge zweier Formanten mit der Bedingung $\omega_p > \sigma_p$ und $\sigma_p > \omega_p$ und ihre Lage in der komplexen Ebene

Schreibt man den Frequenzgang eines Formanten nach Gl.(86) und den eines Antiformanten nach Gl.(87), erkennt

$$G_{dB}(\omega) = 20 \cdot lg \left| \frac{s_p \cdot s_p^*}{(s-s_p)\cdot(s-s_p^*)} \right| \qquad (86)$$

$$H_{dB}(\omega) = -20 \cdot lg \left| \frac{s_z \cdot s_z^*}{(s-s_z)\cdot(s-s_z^*)} \right| \qquad (87)$$

man die Uebereinstimmung bis auf das Vorzeichen. Daraus ist ersichtlich, dass die folgenden Ueberlegungen, die am Beispiel des Formanten erlaeutert werden, ebenso fuer einen Antiformanten gelten.

Die Abb.65 zeigt die Darstellung zweier Formanten, sowohl ihre Lage in der komplexen Ebene, als auch den Frequenzgang. Zum Verlauf von G_{dB2} gehoert das Polpaar s_{p2}, s_{p2}^* fuer das die Bedingung $\omega_p < \sigma_p$ gilt. G_{dB1} ist charakteristisch fuer ein Polpaar s_{p1}, s_{p1}^*, bei dem $\omega_p > \sigma_p$ ist. Da der Fall $\omega_p > \sigma_p$ in der Praxis der wichtigste Fall ist, soll er hier ausschliesslich behandelt werden.

Berechnung erster Schaetzwerte

Der Verlauf von $G_{dB1}(\omega)$ weist ein Maximum bei der Kreisfrequenz

$$\omega_m = \sqrt{\omega_1^2 - \sigma_1^2} \qquad (88)$$

auf. Fuer $\omega_1 >> \sigma_1$ gilt $\omega_m \approx \omega_1$.
In Abb.65 ist die Bandbreite

$$B = \frac{\omega_o - \omega_u}{2\pi} \qquad (89)$$

eingefuehrt worden. Die Kreisfrequenzen ω_u und ω_o sind dabei die Frequenzen, bei denen der Frequenzgang vom Maximum um 3 dB abgefallen ist. Unter der Voraussetzung $\omega_1 >> \sigma_1$ ergibt sich rechnerisch:

$$\omega_o - \omega_u = 2\sigma_1 = 2\pi B \rightsquigarrow \sigma_1 = \pi \cdot B \qquad (90)$$

Bei der Analyse der Uebertragungsfunktion des Vokaltraktes, die nach Gl.(85) als Summe von Formanten und Antiformanten dargestellt werden kann, wird zunaechst in dem Intervall ein Formant gesucht, in dem ein fuer Formanten typischer Verlauf (siehe G_{dB1} in Abb.65) vorliegt, der zu beiden Seiten des Maximums einen Abfall von mindestens 3 dB aufweist. Das ist i.a. dort der Fall, wo $\omega_p >> \sigma_p$ und wo die benachbarten Formanten bzw. Antiformanten weit genug voneinander entfernt sind. Unter dieser Bedingung koennen fuer die Formanten die folgenden Schaetzwerte ermittelt werden:

$$\omega_2 \approx \omega_m \qquad (91)$$
$$\sigma_2 \approx \pi \cdot B$$

Die Abb.66 zeigt noch einmal die entsprechende Darstellung zu Abb.65 fuer einen Antiformanten.

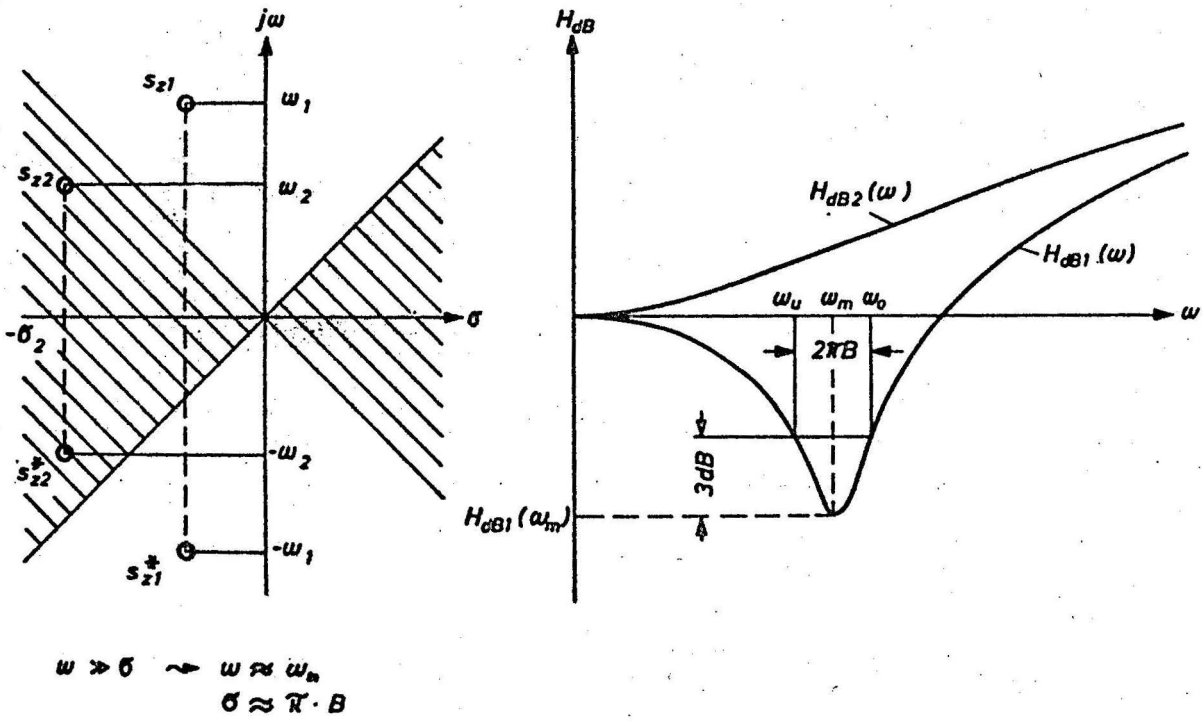


Abb.66, Frequenzgaenge zweier Antiformanten mit der Bedingung $\omega_z > \sigma_z$ und $\sigma_z > \omega_z$ und ihre Lage in der komplexen Ebene

Auch hier ergeben sich die Schaetzwerte:

$$\omega_2 \approx \omega_m \qquad \sigma_2 = \pi \cdot B$$

Test der Schaetzwerte

Fuer die folgenden Betrachtungen wird zunaechst einmal angenommen, dass die Uebertragungsfunktion T(s) nur einen Formanten enthaelt. Der Formant habe die Pole s_{p1} und s_{p1}^* . Die ermittelten Schaetzwerte ω_2 und σ_2 fuer die Polkreisfrequenz und die Daempfung seien ungenau. Es sei zwar $\omega_2 = \omega_1$, aber $\sigma_2 > \sigma_1$. Dann ergibt sich statt des analysierten Frequenzgangs G_{dB1} der Verlauf von G_{dB2} bzw. die Differenzkurve $G_{dB1} - G_{dB2}$ in Abb.67. Die Differenzkurve, die ein Maximum bei ω_m aufweist, und bei der der Betrag des Minimums bei 0 dB liegt, ist ihrerseits charakteristisch dafuer, dass zwar die Polkreisfrequenz ω_2 richtig, die Daempfung σ_2 aber zu gross bestimmt wurde.

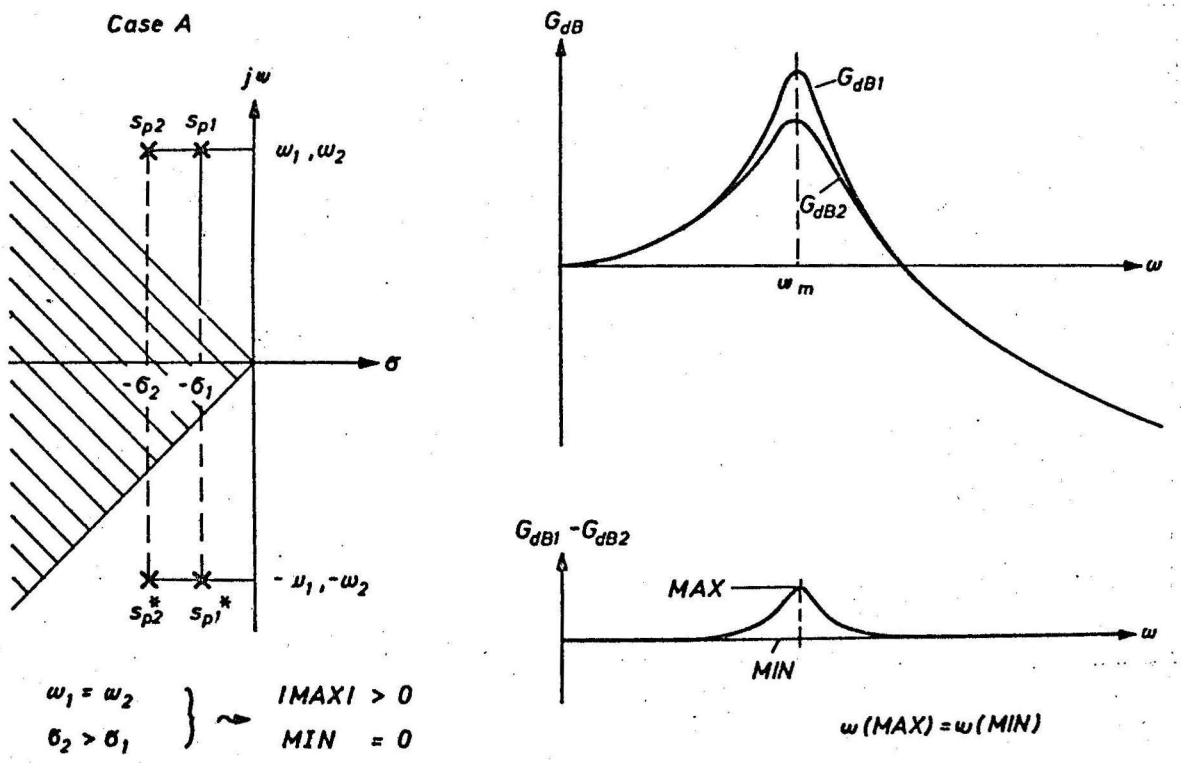


Abb.67, Frequenzgaenge zweier Formanten mit $\omega_1 = \omega_2$ und $\sigma_2 > \sigma_1$ sowie der Verlauf der Differenzkurve

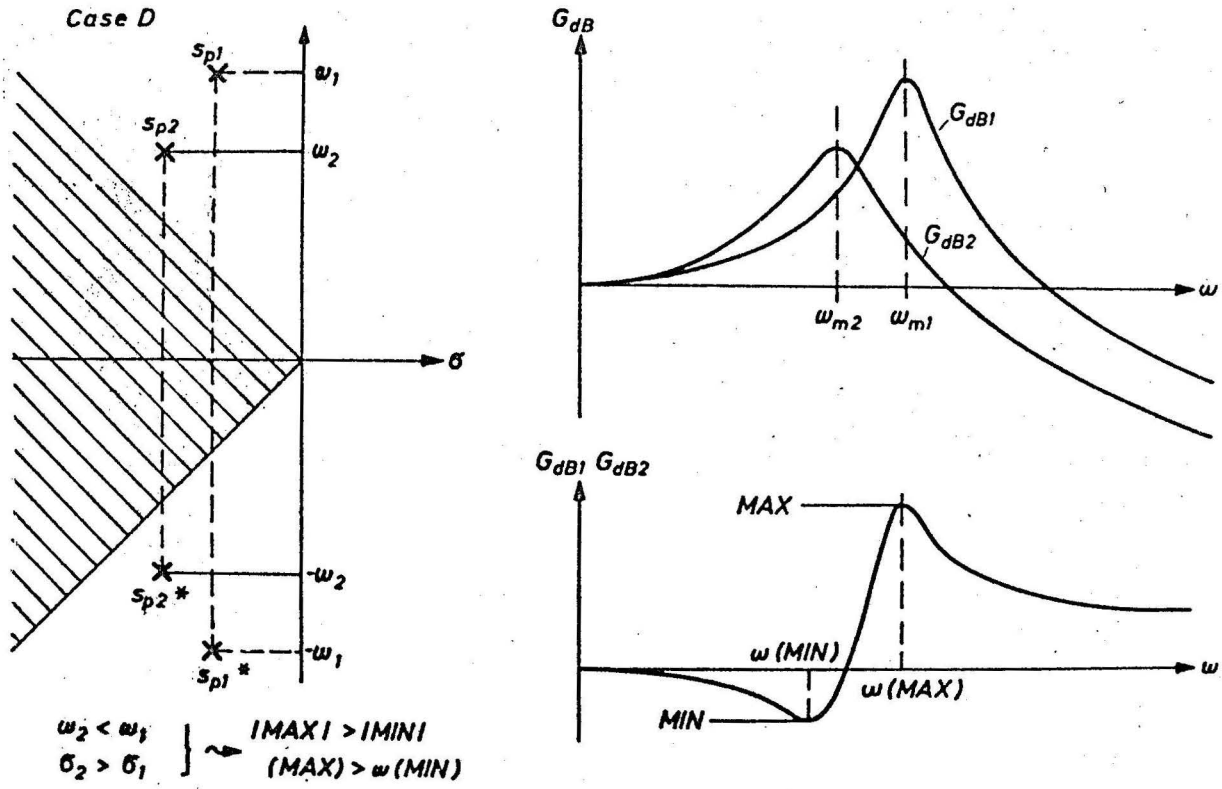


Abb.68, Frequenzgaenge zweier Formanten mit $\omega_2 < \omega_1$ und $\sigma_2 > \sigma_1$ sowie der Verlauf der Differenzkurve

Abb.68 zeigt einen komplizierteren Fall. Der vorgegebene Frequenzgang, der analysiert werden soll, ist G_{dB1}. Die Abschaetzung der Lage der Pole s_{p1} und s_{p1}^* fuehrt nach Gl.(91) zu den falschen Werten s_{p2} und s_{p2}^*. Berechnet man die Differenz des zu analysierenden und des aus den Schaetzwerten berechneten Frequenzgangs, so erhaelt man eine Kurve, die die folgenden Merkmale aufweist: Der Betrag des Maximums ist groesser als der Betrag des Minimums, und der zum Maximum gehoerende Frequenzwert liegt bei hoeheren Werten, als der zum Minimum gehoerende. Man kann jetzt aus diesem charakteristischen Verlauf der Differenzkurve zurueckschliessen, dass die Polkreisfrequenz ω_2 zu klein und die Daempfung σ_2 zu gross bestimmt wurden.

Alle moeglichen charakteristischen Faelle fuer den Verlauf der Differenzkurve sind in der Tabelle nach Abb.69 zu-

CASE				
A	IMAXI>0 MIN=0	ω (MAX) = ω (MIN)	$\omega_2 = \omega_1$	$\sigma_2 > \sigma_1$
B	MAX=0 IMINI>0	ω (MAX) = ω (MIN)	$\omega_2 = \omega_1$	$\sigma_2 < \sigma_1$
C	IMAXI > IMINI	ω (MAX) < ω (MIN)	$\omega_2 > \omega_1$	$\sigma_2 > \sigma_1$
D	IMAXI > IMINI	ω (MAX) > ω (MIN)	$\omega_2 < \omega_1$	$\sigma_2 > \sigma_1$
E	IMAXI < IMINI	ω (MAX) < ω (MIN)	$\omega_2 > \omega_1$	$\sigma_2 < \sigma_1$
F	IMAXI < IMINI	ω (MAX) > ω (MIN)	$\omega_2 < \omega_1$	$\sigma_2 < \sigma_1$
G	IMAXI = IMINI	ω (MAX) < ω (MIN)	$\omega_2 > \omega_1$	$\sigma_2 = \sigma_1$
H	IMAXI = IMINI	ω (MAX) > ω (MIN)	$\omega_2 < \omega_1$	$\sigma_2 = \sigma_1$

Abb.69, Zur Charakterisierung aller moeglichen Differenzkurven

sammengestellt und mit den Buchstaben A bis H gekennzeichnet. Mit Hilfe dieser Tabelle laesst sich jetzt ein Algorithmus herleiten, der sagt, in welchem Sinn die Schaetzwerte ω_2 und σ_2 variiert werden muessen, um ein besseres Ergebnis zu erzielen.

In Abb.70 ist ein Teil des Iterationsschemas dargestellt, innerhalb dessen die fehlerhaft bestimmten Formantfrequenz und Bandbreitewerte variiert werden.

Iteration

Die einzelnen Teile des Iterationsschemas sind durch Positionsnummern gekennzeichnet. Die Positionsnummern laufen von

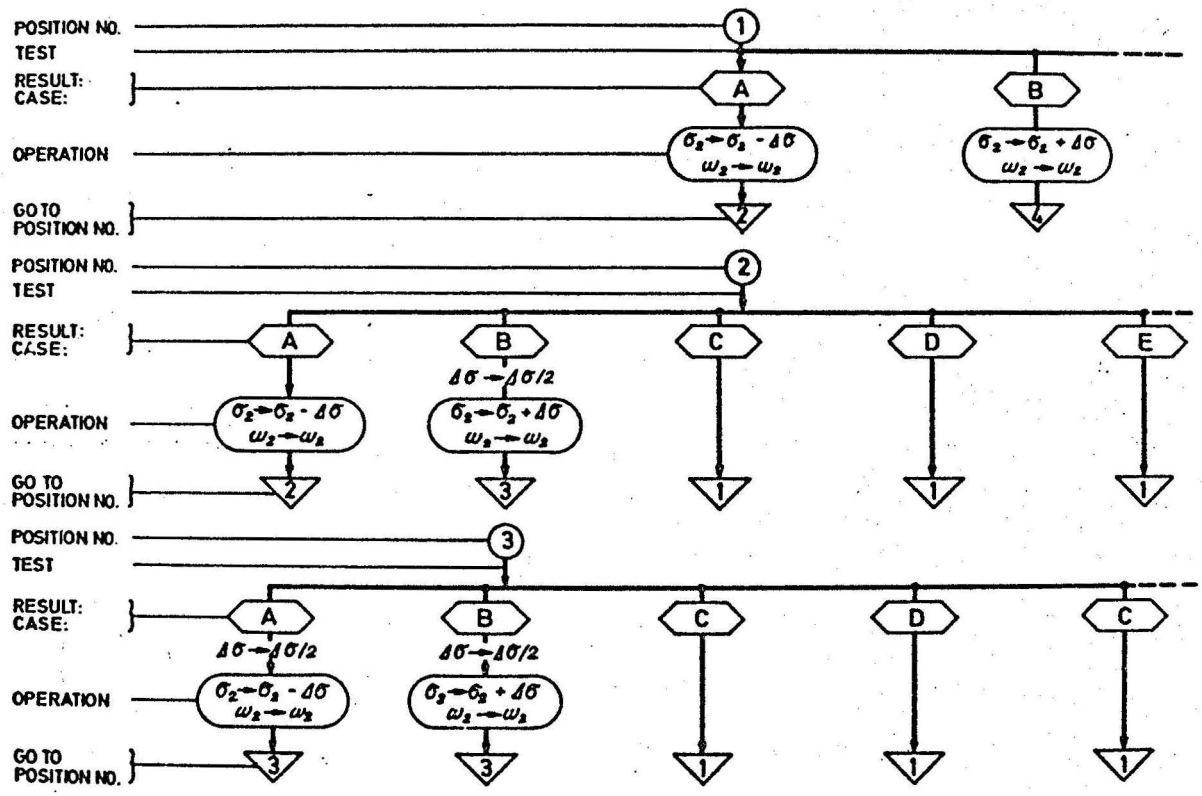


Abb.70, Ausschnitt aus dem Iterationsschema zur Korrektur der Schaetzwerte

1 bis 16 und geben einen Hinweis darauf, welche Iterationsschritte fuer den zu analysierenden Formanten bereits durchgefuehrt worden sind.

Der Ausgangspunkt fuer das Iterationsschema ist die Positionsnummer 1. Hier liegen bereits die Schaetzwerte nach Gl.(91) vor. Es wird zunaechst mit diesen Schaetzwerten die Differenzkurve berechnet und aus der Tabelle nach Abb.69 der Fall herausgesucht, fuer den die Differenzkurve charakteristisch ist. Dieser Vorgang wird in Abb.70 mit 'Test' bezeichnet. Das Ergebnis des Tests ist z.B. Fall A. Aus der Tabelle nach Abb.69 ergibt sich, wie die Frequenz- und Daempfungswerte geaendert werden muessen. Die Schaetzwerte werden nach der Vorschrift aus Abb.69 variiert und ein neuer Zustand des Iterationsprozesses wird durch die neue Positionsnummer 2 angezeigt. Anschliessend wird mit den verbesserten Werten erneut der Test durchgefuehrt und die Schaetzwerte weiter verbessert, usw. Die Iteration wird abgebrochen:

a) wenn der Schaetzwert kleiner oder gleich dem vorgegebenen Wert 0.0001 Hz ist
b) wenn die Differenz |MAX| - |MIN| kleiner als 0.1 dB ist

c) wenn die Schrittweiten $\Delta\sigma$, $\Delta\omega$ kleiner als 0.1 Hz sind
d) wenn die Anzahl der Iterationen groesser als 20 ist.

Ergebnis

Fuer Genauigkeitsuntersuchungen wurde der Frequenzgang zuerst aus bekannten ω - und σ -Werten berechnet und dann anschliessend wieder analysiert. Es wurde darauf geachtet, dass die oben genannten Bedingungen bezueglich des Abstandes der Formanten bzw. der Antiformanten untereinander und die Bedingung $\omega_p >> \sigma_p$ eingehalten wurden. Unter diesen Voraussetzungen konnten die Frequenzwerte auf weniger als 0.5% und die Bandbreitewerte auf weniger als 5% ihres wahren Wertes genau bestimmt werden.

Die halbautomatische Formantbestimmung mit dem Display benoetigt durch den Einsatz eines Operateurs sehr viel Rechenzeit. Selbst ein geuebter Operateur benoetigt fuer die Analyse einer Uebertragungsfunktion ca 5 min, d.h. fuer die Analyse von 1 sek Sprache ca 500 min.

Das Verfahren eignet sich besonders gut zum Studium einzelner Kurzzeitspektren bzw. Uebertragungsfunktionen des Vokaltraktes und zur Untersuchung oder Entwicklung neuer Formantbestimmungsverfahren. Ausserdem wurde es vom Verfasser dazu benutzt, fuer das Momentverfahren nach NAKATSIN und SUZUKI an kritischen Stellen, z.B. den Uebergaengen von stimmlosen zu stimmhaften Lauten, neue Anfangswerte zu finden.

In einer weiteren Arbeit, die von VORMELCHER /43/ ausgefuehrt wurde, ist das oben beschriebene Verfahren automatisiert worden, d.h. die Aufgabe des Operateurs wurde von einem Programm wahrgenommen. Es werden dabei drei Formanten und, falls vorhanden, eine Nullstelle nach Frequenz und Bandbreite aus der Uebertragungsfunktion herausgesucht. Da das Verfahren insgesamt sehr aufwendig ist, werden pro 10ms-Intervall der Sprache ca 1 min Rechenzeit fuer die Formantbestimmung benoetigt. Das entspricht 100 min Rechenzeit bei der Analyse von 1 sek Sprache.

6.4 Formantbestimmung im Zeitbereich

6.4.1 Methoden der Formantbestimmung

Es gibt im wesentlichen zwei Moeglichkeiten, wie man unmittelbar aus einer Sprachzeitfunktion, in deren Kurzzeitspektrum ein dominierender Formant vorliegt, die Formantfrequenz bestimmen kann:

1. aus dem mittleren Abstand der Nulldurchgaenge
2. aus dem mittleren Abstand der Maxima und Minima

Die Formantbestimmung aus dem mittleren Abstand der Maxima und Minima entspricht der Formantbestimmung aus dem mittleren Abstand der Nulldurchgaenge, wenn man die Zeitfunktion vorher differenziert. Das bedeutet, dass durch die Verwendung der Maxima und Minima Resultate erzielt werden, die einer Hoehenanhebung der Sprache entsprechen. Man kann diese Tatsache dann besonders vorteilhaft ausnutzen, wenn in der Sprache zwei etwa gleichwertige Formanten enthalten sind und der Frequenzwert des hoeheren Formanten ermittelt werden soll.

6.4.2 Formantbestimmung durch Bandpassfilterung

In 6.3.3 werden von SCHAFER und RABINER fuer maennliche Sprecher die folgenden Frequenzbereiche fuer die ersten drei Formanten angegeben:

1. Formant: 200 Hz bis 900 Hz
2. Formant: 550 Hz bis 2700 Hz
3. Formant: 1100 Hz bis 2950 Hz

Daraus ist ersichtlich, dass die geringste Ueberlappung der Frequenzbereiche zwischen dem ersten und dem zweiten Formanten auftritt. Dagegen sind der zweite und der dritte Formant aufgrund ihrer Frequenzbereiche praktisch nicht voneinander zu trennen.

Der erste Formant wird mit einem Bandpass, der einen Durchlassbereich von 200 Hz bis 1000 Hz aufweist, aus der Zeitfunktion herausgefiltert. Der Frequenzwert des ersten Formanten wird aus den Nulldurchgaengen am Ausgang des Bandpass berechnet.
Der zweite und dritte Formant werden mit einem zweiten Bandpass herausgefiltert, der einen Durchlassbereich von 900 Hz bis 2700 Hz aufweist. Da der zweite Formant im Kurzzeitspektrum i.a. eine etwas groessere Amplitude als der dritte Formant aufweist, kann man aus den Nulldurchgaengen am Ausgang des zweiten Bandpasses naeherungsweise den Frequenzwert des zweiten Formanten bestimmen. Der dritte Formant kann wegen der schon erwaehnten Ueberlappung der Bereiche nicht mehr durch eine weitere Bandpassfilterung ermittelt werden. Da der dritte Formant ausserdem fuer die Verstaendlichkeit der Sprache eine geringe Bedeutung hat, wurde er gleich dem konstanten Wert von 2500 Hz angesetzt.

Die verwendeten Bandpaesse wurden in Frequency- Sampling-Technique programmiert und benoetigen daher sehr viel Rechenzeit. Fuer die Verarbeitung von 1 sek Sprache wurde eine Rechenzeit von 1386 sek gemessen.

6.4.3 Formantbestimmung durch inverse Filterung

Aus der Gl.(19) kann mit Gl.(12) die folgende Beziehung fuer das Druckspektrum im Schallfeld eines sprechenden Menschen abgeleitet werden:

$$P_1(s) = Q(s) \cdot R(s) \cdot \prod_{n=1}^{\infty} H_n(s) \qquad (92)$$

Geht man davon aus, dass die Uebertragungsfunktion des Vokaltraktes im wesentlichen durch drei Formanten dargestellt wird, und fasst man die Abstrahlung und die hoeheren Formanten in dem Faktor K(s) zusammen, erhaelt man:

$$P_1(s) = H_1(s) \cdot H_2(s) \cdot H_3(s) \cdot Q(s) \cdot K(s) \qquad (93)$$

Da K(s) zeitlich konstant ist, die anderen Groessen aber in ihrer Abhaengigkeit von der Zeit bestimmt werden sollen, kann man die Kurzzeitspektren einfuehren und erhaelt dann:

$$P_1(s,t) = H_1(s,t) \cdot H_2(s,t) \cdot H_3(s,t) \cdot Q(s,t) \cdot K(s) \qquad (94)$$

Die zugehoerige Zeitfunktion heisse $p_1(t)$.

Den dominierenden Anteil im Spektrum $P_1(s,t)$ stellt i.a. der erste Formant dar, insbesondere dann, wenn man die Zeitfunktion $p_1(t)$ mit einem Tiefpass der Grenzfrequenz von 1 kHz filtert und so die hoeheren Frequenzanteile der Sprache abschwaecht. Das durch einen Tiefpass gefilterte Signal $p_1(t)$ soll $p_{1T}(t)$ genannt werden. Aus dem mittleren Abstand der Nulldurchgaenge der Zeitfunktion $p_{1T}(t)$ kann dann der Frequenzwert des ersten Formanten berechnet werden. Der zugehoerige Daempfungswert wird aus einer Tabelle berechnet, die von FLANAGAN (/2/ S.152) angegeben wird.

Durch inverse Filterung der Zeitfunktion $p_1(t)$, d.h. durch Filterung mit einem Antiformanten der Nullstellenfrequenz f1 und der Daempfung σ_1 wird der erste Formant aus der Zeitfunktion herausgefiltert. Das Kurzzeitspektrum berechnet sich dann zu:

$$P_2(s,t) = H_2(s,t) \cdot H_3(s,t) \cdot Q(s,t) \cdot K(s) \qquad (95)$$

Die zugehoerige Zeitfunktion sei $p_2(t)$.

Die dominierenden Formanten sind jetzt der zweite und der dritte Formant. Die Einfluesse der hoeheren Formanten werden durch Filterung mit einem Tiefpass der Grenzfrequenz von 3300 Hz eliminiert und man erhaelt $p_{2T}(t)$. Da der zweite und dritte Formant in den meisten Faellen einander gleichwertig sind, wird aus der Zeitfunktion $p_{2T}(t)$ zuerst der dritte Formant f3 aus dem mittleren Abstand der Maxima und Minima berechnet. Der zugehoerige Daempfungswert σ_3 ergibt sich wiederum aus der oben genannten Tabelle.

Die urspruengliche Zeitfunktion, aus der bereits der erste Formant herausgefiltert worden war, wird jetzt mit einem Antiformant der Frequenz f3 und der Daempfung σ_3 gefil-

tert. Die Beschreibung des Kurzzeitspektrums reduziert sich damit nach Gl.(96) zu:

$$P_3(s,t) = H_2(s,t) \cdot Q(s,t) \cdot K(s) \qquad (96)$$

Die zugehoerige Zeitfunktion heisst $p_3(t)$.

Die Zeitfunktion $p_3(t)$ wird jetzt mit einem Tiefpass der Grenzfrequenz von 2700 Hz gefiltert, um die hoeheren Formanten zu entfernen. Man erhaelt dabei die Zeitfunktion $p_{3T}(t)$. Im Kurzzeitspektrum ist dann der zweite Formant der dominierende. Der Frequenzwert f2 wird aus dem mittleren Nullpunktsabstand von $p_{3T}(t)$ berechnet und die Daempfung σ_3 der Tabelle entnommen.

Den Aufbau des verwendeten Programms zur Formantextraktion durch inverse Filterung der Zeitfunktion ist in Abb.71 dargestellt:

Die verwendeten Sprachbeispiele wiesen einen hohen Brummanteil auf, der sich besonders nachteilig bei der Frequenzbestimmung aus den mittleren Nullpunktsabstaenden bemerkbar machte. Die Sprache muss deshalb zuerst durch einen Hochpass gefiltert werden. Der Hochpass besteht aus einem Formanten der Polfrequenz von 200 Hz und der Daempfung von 90 Hz und einem Antiformanten der Nullstellenfrequenz von 45 Hz und der Daempfung von 1 Hz.

Die Formantanalyse wird in 10 ms-Schritten entsprechend einer Schrittzahl von jeweils 100 Abtastwerten durchgefuehrt. Die 100 Abtastwerte, die durch den Hochpass gefiltert wurden, werden als $p_1(t)$ abgespeichert. Die Werte $p_1(t)$ werden durch einen Tiefpass der Grenzfrequenz von f_g =1 kHz gefiltert und als $p_{1T}(t)$ abgespeichert. Aus dem mittleren Nullpunktsabstand der Werte $p_{1T}(t)$ wird der erste Formant und aus der genannten Tabelle der zugehoerige Daempfungswert berechnet. Die gepeicherten Funktionswerte $p_1(t)$ koennen jetzt mit einem Antiformanten gefiltert werden, der die Nullstellenfrequenz f1 und die Daempfung σ_1 hat. Die gefilterten Werte werden als $p_1(t)$ abgespeichert. Anschliessend werden die Werte $p_2(t)$ durch den Tiefpass gefiltert, der jetzt die Grenzfrequenz f_g =3.3 kHz hat. Aus der gefilterten Zeitfunktion $p_{2T}(t)$ wird der mittlere Abstand der Extremwerte bestimmt und daraus f3 und ueber die Tabelle σ_3 berechnet. Die noch gespeicherte Zeitfunktion $p_2(t)$ kann jetzt durch einen Antiformanten der Nullstellenfrequenz f3 und der Daempfung σ_3 gefiltert werden und man erhaelt $p_3(t)$.

Die Funktion $p_3(t)$ wird mit einem Tiefpass der Grenzfrequenz f_g =2.7 kHz gefiltert und aus dem mittleren Nullpunktsabstand f2 und σ_2 ermittelt.

Nach einmaligem Durchlauf des Programms entsprechend dem Blockschaltbild nach Abb.71 wird auf die gleiche Weise das naechste 10 ms-Intervall analysiert.

Die Rechenzeit betraegt fuer die Analyse von 1 sek Sprache auf dem Rechner CAE 90-40 ca 300 sek.

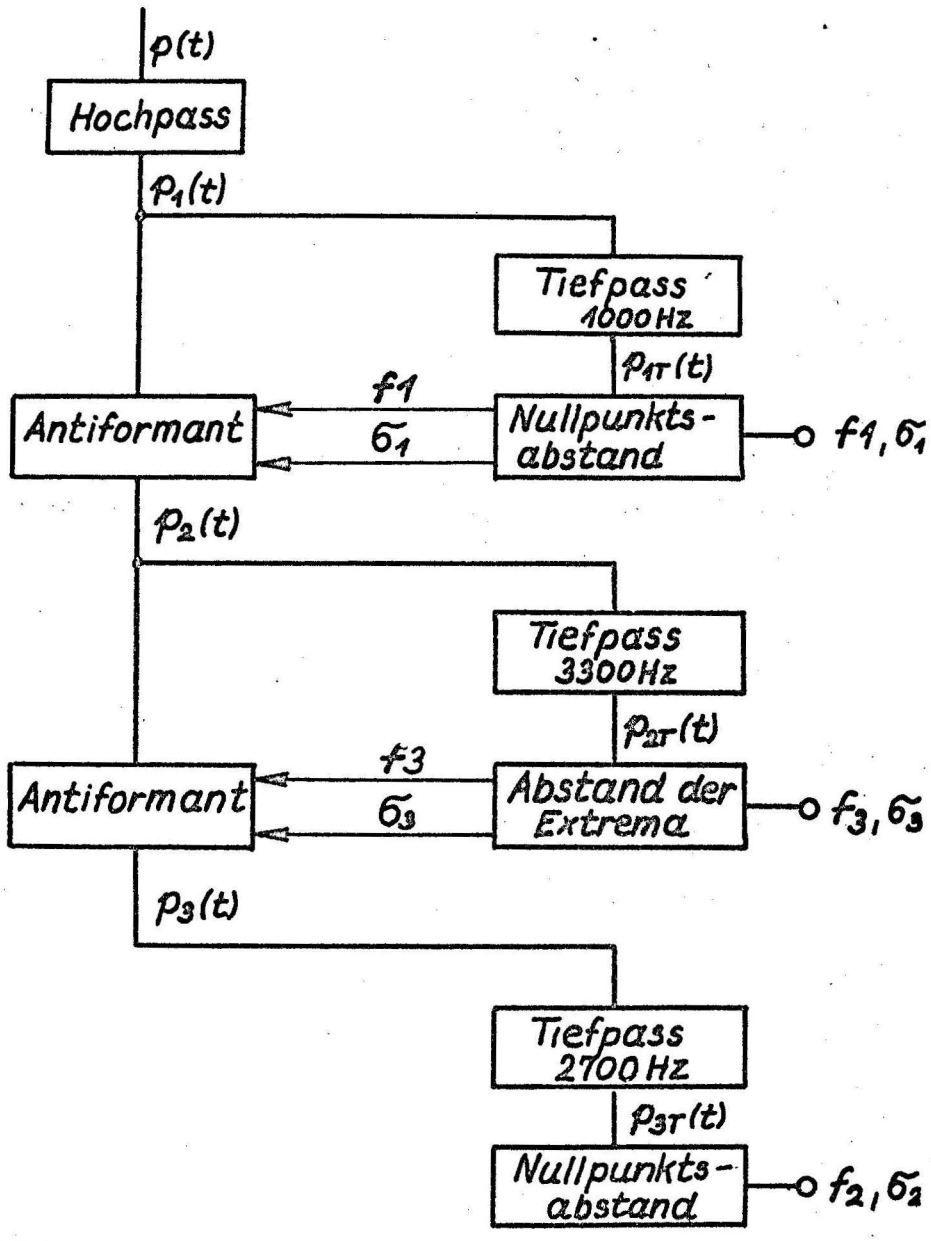


Abb.71, Blockschaltbild des Programms zur Formantbestimmung mittels inverser Filterung der Zeitfunktion

6.5 Vergleich der Formantbestimmungsverfahren

Die in 6.3 und 6.4 genannten Formantbestimmungsverfahren wurden zur Analyse der Worte 'HAWAII', 'SCHEUSSLICH', 'BODEN' und 'LEGEN-MOECHTE' verwendet. Die ermittelten Formantverlaeufe fuer die Beispiele 'HAWAII' und 'LEGEN-MOECHTE' sind graphisch in Abb.72 bis Abb.77 dargestellt. In den einzelnen Abbildungen sind der erste, zweite und dritte Formant uebereinander und mit demselben Frequenzmassstab dargestellt. Die senkrechten Linien geben die Abgrenzungen der stimmhaften und stimmlosen Bereiche in den einzelnen Worten an. Die stimmhaften Bereiche sind unmittelbar ueber dem Zeitmassstab mit einer '1' und die stimmlosen Bereiche mit einer '0' bezeichnet.

Aus den dargestellten Formantverlaeufen wurde mit einem Synthetisator nach Abb.48 die Sprache synthetisiert. Auf diese Weise konnten die Formantbestimmungsverfahren auch akustisch miteinander verglichen werden. Bei der Formantbestimmung aus den spektralen Momenten nach 6.3.2 sind feste, sich nicht ueberlappende Frequenzbereiche zur Momentbildung angenommen worden. Ein Vergleich von der Abb.72 mit Abb.77 zeigt, dass der Verlauf der ersten Formanten, dessen Frequenzbereich sich kaum mit dem des zweiten Formanten ueberlappt, sehr gut wiedergegeben wird. Der Verlauf des zweiten Formanten wird dagegen zu hohen Frequenzen hin abgeschnitten, wie es besonders am charakteristischen Verlauf des 'ai' in 'HAWAII' aus Abb.72a im Vergleich mit Abb.77a ersichtlich ist. Das 'HAWAII' nach 6.3.2 ist deshalb sehr schlecht verstaendlich im Gegensatz zum 'HAWAII' nach 6.4.3.
Die Verlaeufe des dritten Formanten unterscheiden sich bei allen Formantbestimmungsverfahren sehr stark. Da ihr Einfluss auf die Guete der synthetisierten Sprache nur sehr gering ist, soll der Verlauf des dritten Formanten hier und bei allen anderen Formantbestimmungsverfahren nicht weiter besprochen werden.

Bei dem Momentverfahren nach NAKATSIN und SUZUKI, das in Kap 6.3.3 behandelt wurde, sind im Gegensatz zu 6.3.2 variable Grenzen zur Momentbildung angenommen worden. Beim Vergleich der Abb.73a mit Abb.72a zeigt es sich, dass der Verlauf des zweiten Formanten in der Naehe des 'ai' von 'HAWAII' sehr viel besser wiedergegeben ist, als im vorangegangenen Beispiel.

Aus der Verwendung variabler Grenzen kann sich jedoch auch ein grosser Nachteil ergeben, wie er im gleichen Beispiel zu finden ist. Der Verlauf des ersten Formanten hat sich von Anfang an in die falsche Richtung bewegt und so wurden aus falschen Anfangswerten immer neue falsche Werte berechnet, so dass der Verlauf des ersten Formanten in der Abb.73a voellig unbrauchbar und das Wort 'HAWAII' akustisch unverstaendlich geworden ist. Im Gegensatz dazu hoert sich das Sprachbeispiel 'LEGEN-MOECHTE' sehr gut an.

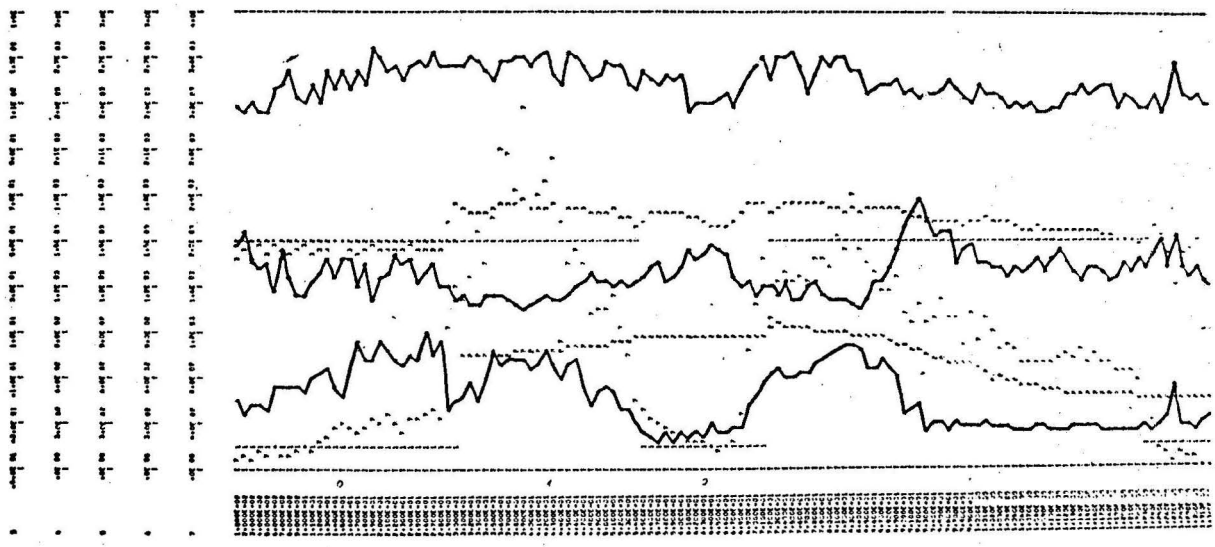


Abb.72a, 'HAWAII', Formantbest. nach Kap 6.3.2

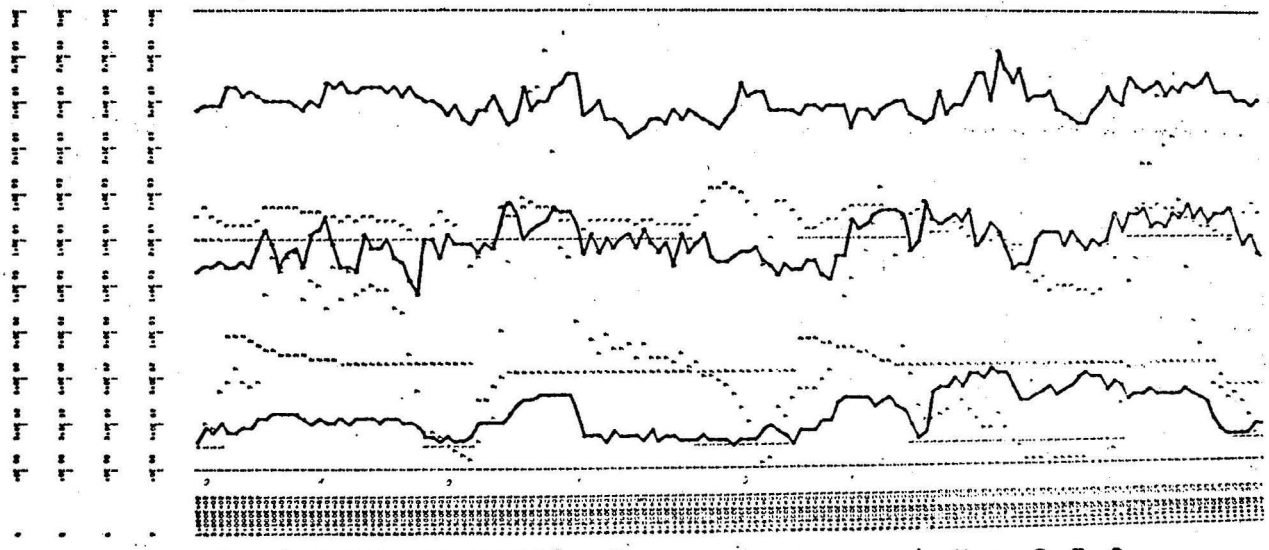


Abb.72b, 'LEGEN-MOECHTE', Formantbest. nach Kap 6.3.2

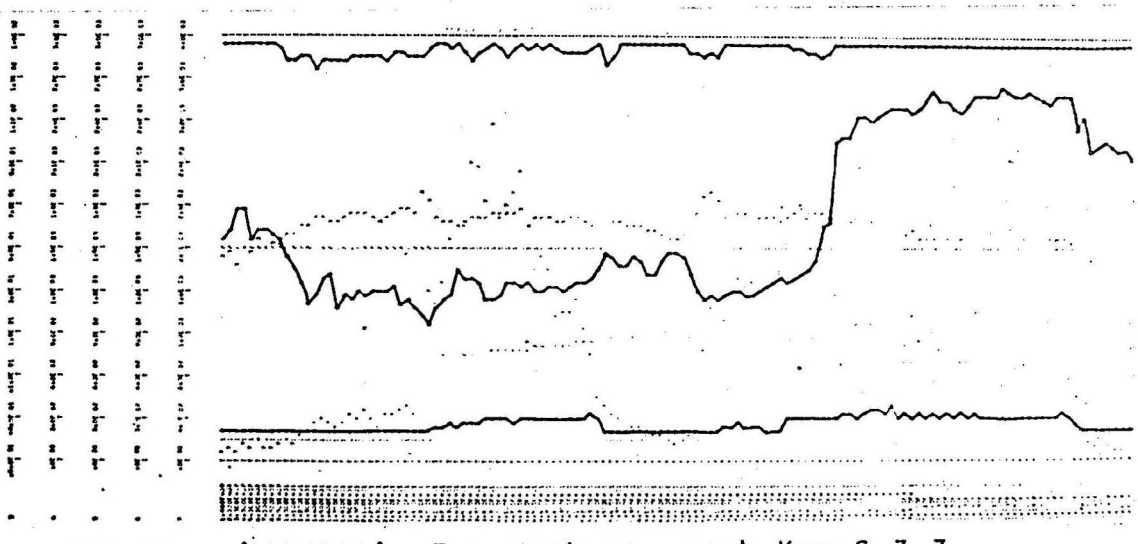


Abb.73a, 'HAWAII', Formantbest. nach Kap 6.3.3

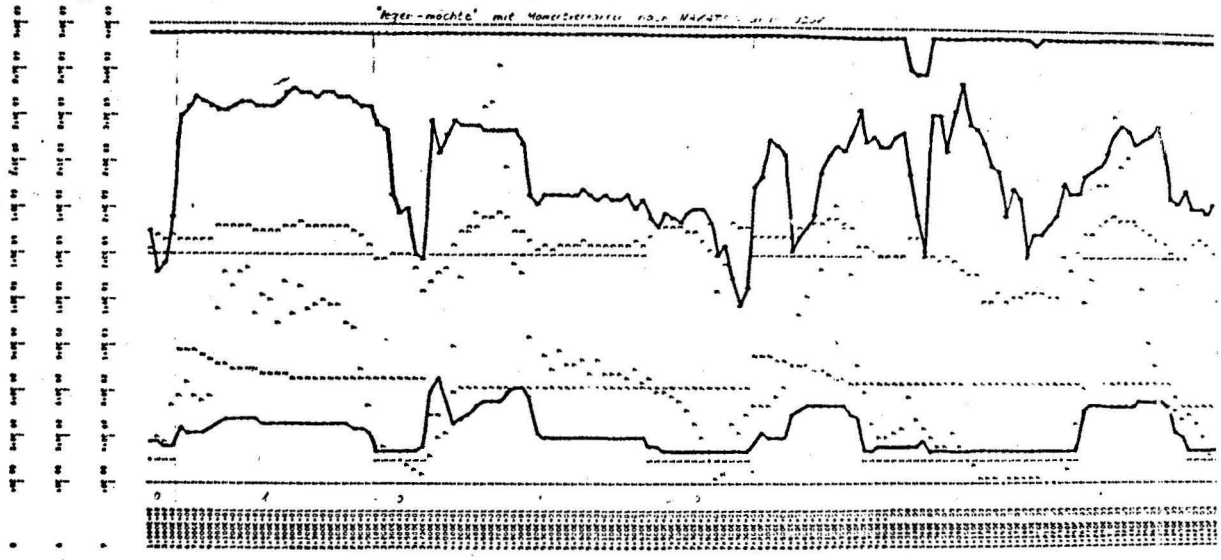


Abb.73b, 'LEGEN-MOECHTE', Formantbest. nach Kap 6.3.3

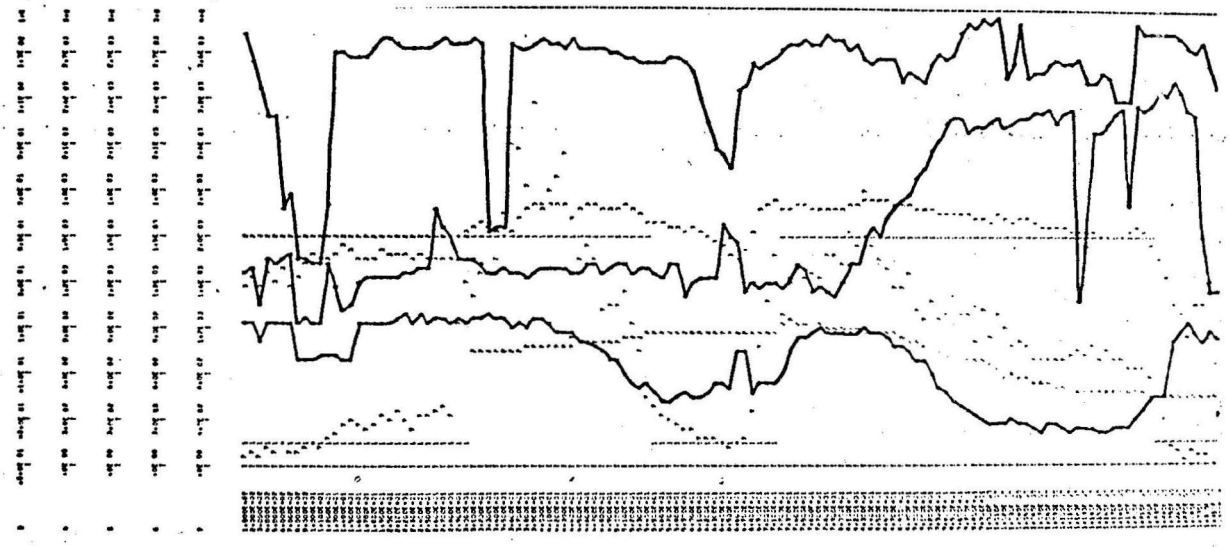


Abb.74a, 'HAWAII', Formantbest. nach Kap 6.3.4

Abb.74b, 'LEGEN-MOECHTE', Formantbest. nach Kap 6.3.4

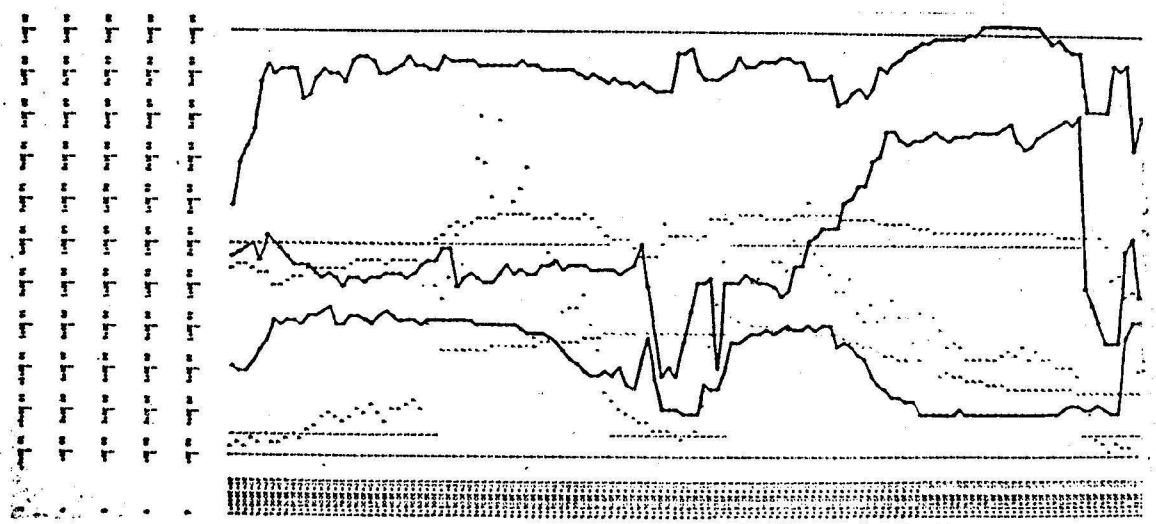


Abb.75a, 'HAWAII', Formantbest. nach Kap 6.3.5

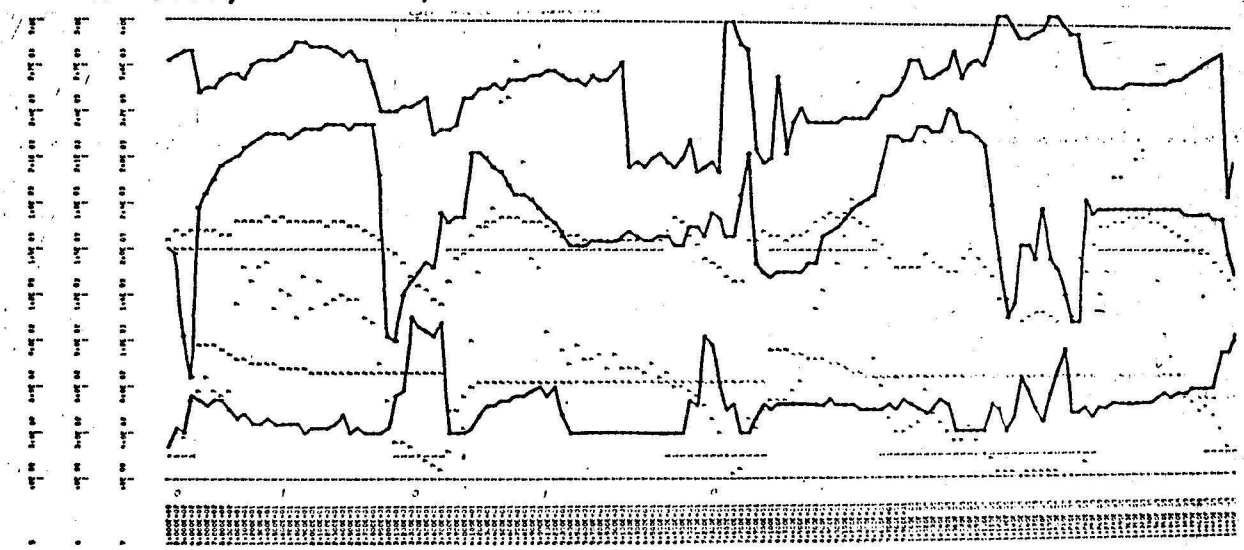


Abb.75b, 'LEGEN-MOECHTE', Formantbest. nach Kap 6.3.5

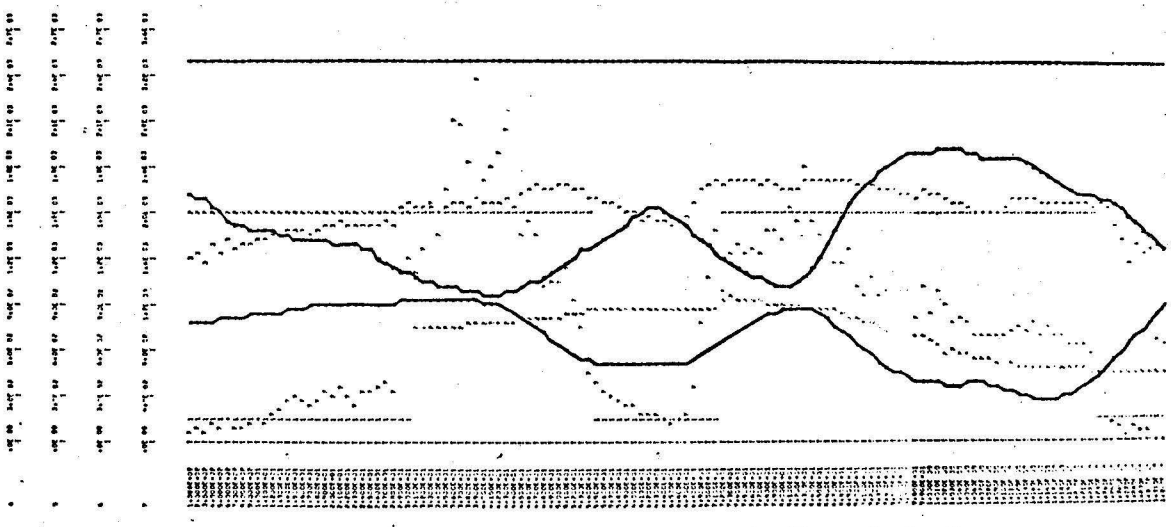


Abb.76a, 'HAWAII', Formantbest. nach Kap 6.4.2

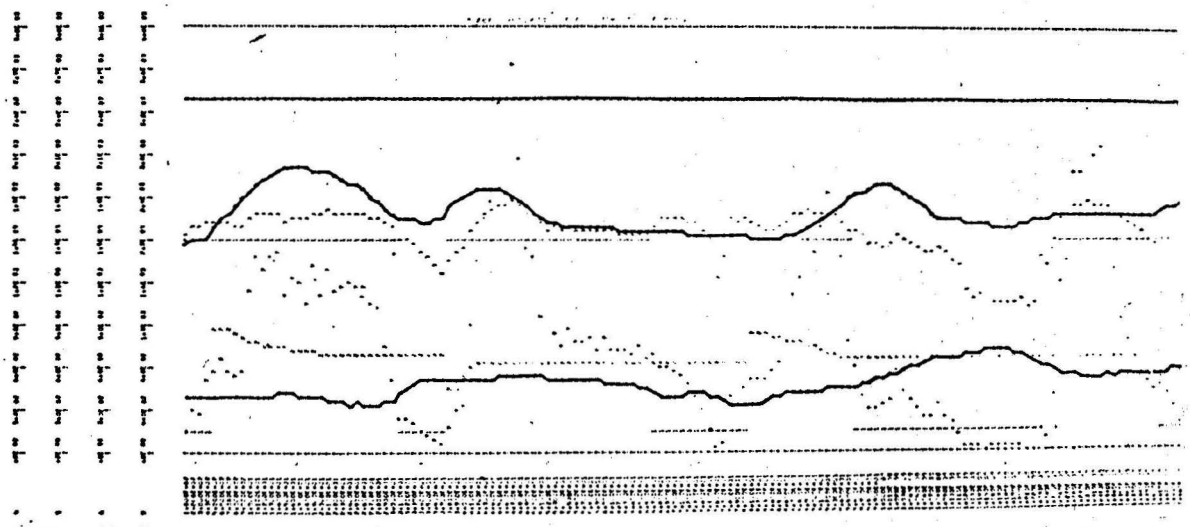


Abb.76b, 'LEGEN-MOECHTE', Formantbest. nach Kap 6.4.2

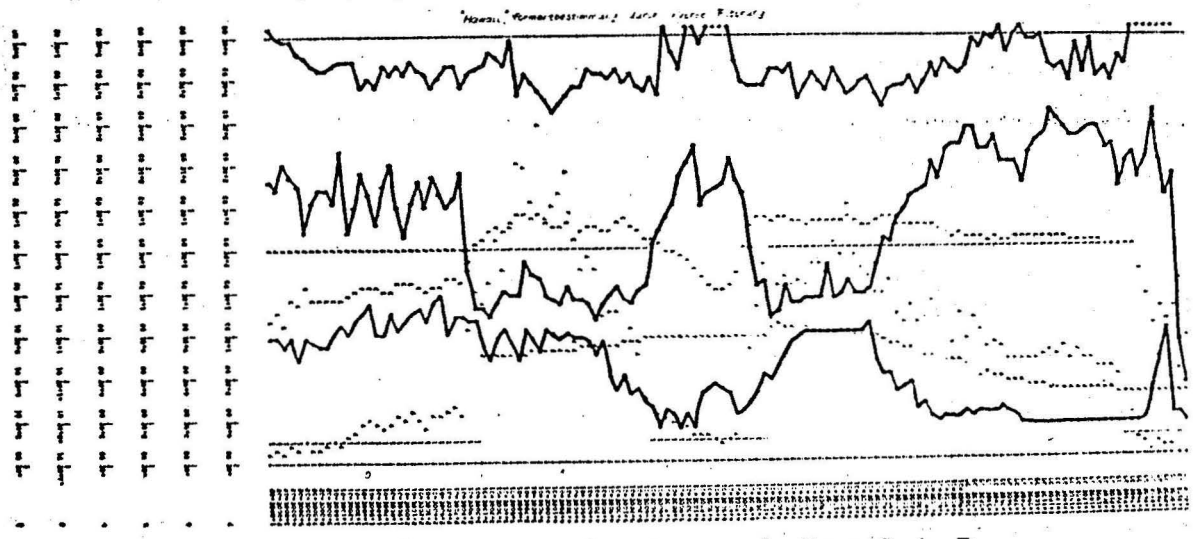


Abb.77a, 'HAWAII', Formantbest. nach Kap 6.4.3

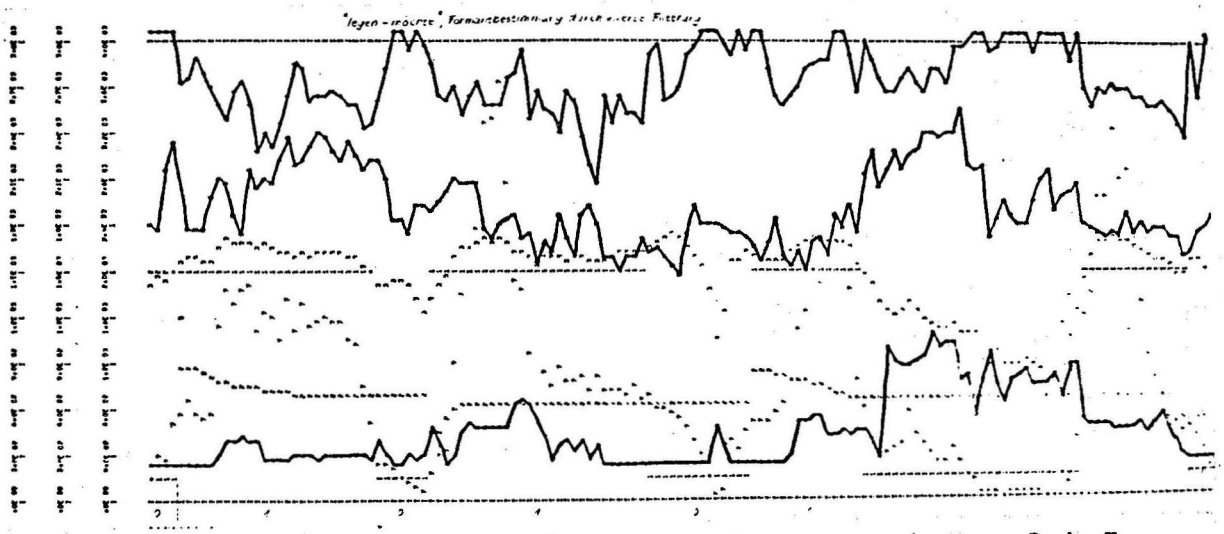


Abb.77b, 'LEGEN-MOECHTE', Formantbest. nach Kap 6.4.3

Bei dem in Kap 6.3.4 behandelten Formantbestimmungsverfahren nach SCHAFER und RABINER werden die Formanten der einzelnen Samples weitgehend unabhaengig voneinander bestimmt. Die Formantverlaeufe werden im wesentlichen durch kontinuierliche und glatte Kurven wiedergegeben, obwohl an manchen Stellen sehr grosse Spruenge in den Frequenzwerten vorkommen. Das Auftreten der grossen Spruenge findet man vor allem in den stimmlosen Bereichen, in denen ohnehin keine Formantstruktur der Sprache zu erwarten ist.

Dasselbe gilt auch fuer das Formantbestimmungsverfahren nach VORMELCHER, das aus dem in Kap 6.3.5 beschriebenen Verfahren hervorgegangen ist.

Die Formantverlaeufe nach Abb.76 wurden durch Bandpassfilterung der Zeitfunktion und Frequenzbestimmung aus dem mittleren Nullpunktsabstand gewonnen. Durch eine nichtlineare Glaettung der 'Ausreisser' wurde der auffaellig glatte Formantverlauf in der Abb.76 erzeugt. Waehrend das Beispiel 'HAWAII' akustisch zu verstehen ist, ist 'LEGEN-MOECHTE' kaum verstaendlich. Der Grund dafuer liegt vor allem in dem flachen Formantverlauf. Es hat sich naemlich bei dem akustischen Vergleich der Formantbestimmungsverfahren gezeigt, dass die Information, die die Sprache verstaendlich macht, nicht etwa in den richtigen Frequenzlagen der Formanten, sondern in der Aenderung, d.h. in den dynamischen Uebergaengen zu hoeheren und tieferen Frequenzbereichen liegt.

Ein weiteres Verfahren, dass sich sehr gut den schnellen Aenderungen der Formantverlaeufe anpassen kann, ist das Verfahren nach Kap 6.4.3, das auf der inversen Filterung der Zeitfunktion beruht. Die Tendenzen der Formantverlaeufe werden einwandfrei wiedergegeben, jedoch sind die Schwankungen im Formantverlauf wesentlich groesser als bei allen anderen Verfahren.

Selbst Formantbestimmungsverfahren, wie das nach SCHAFER und RABINER und das nach VORMELCHER, die bei einem subjektiven Vergleich in ihrer Qualitaet nur unwesentlich voneinander abweichen, weisen stark unterschiedliche Formantverlaeufe in den stimmlosen Bereichen auf. Ein einheitlicher Formantverlauf ist auch nicht zu erwarten, da die Sprache in den stimmlosen Gebieten keine Formantstruktur aufweist. Die geringen Unterschiede bei der subjektiven Beurteilung deuten aber ganz allgemein darauf hin, dass die genaue Lage der Formantverlaeufe in den stimmlosen Teilen der Sprache fuer die Verstaendlichkeit ohne Bedeutung ist.

Beim akustischen Vergleich der Formantbestimmungsverfahren hat es sich gezeigt, dass die einzelnen Verfahren nicht von sich aus gut oder schlecht sind, sondern fuer einzelne Wortbeispiele gute Resultate, fuer andere dagegen wieder sehr schlechte Resultate vorweisen.

In der Tabelle 9 wird vom Verfasser der Versuch unternommen, die Resultate der einzelnen Formantbestimmungsverfahren aufgrund des akustischen Vergleichs mit den Ziffern

1 bis 5 subjektiv zu benoten. Eine sehr hohe Qualitaet soll durch eine '1' und eine sehr geringe Qualitaet durch eine '5' gekennzeichnet werden.

Formant-best.verf. lt.Kap.	HAWAII	SCHEUSS-LICH	BODEN	LEGEN MOECHTE	Mittlere Note
6.3.2	4	4	4	4	4.0
6.3.3	5	3	4	2	3.5
6.3.4	2	3	2	2	2.3
6.3.5	1	2	3	3	2.3
6.4.2	3	4	5	5	4.3
6.4.3	2	3	3	2	2.5

Tabelle 9, Subjektive Beurteilung der Formant-bestimmungsverfahren

Ein weiteres Kriterium fuer die Guete eines Formantbestimmungsverfahrens ist die benoetigte Rechenzeit. In der Tabelle 10 sind deshalb die subjektiv festgelegten Noten den Rechenzeiten fuer die Analyse von 1 sek Sprache gegenuebergestellt.

Formantbest.verf. nach Kapitel	Gesamtnote nach Tab 9	Rechenzeit in sek fuer die Analyse von 1 sek Sprache
6.3.2	4.0	180
6.3.3	3.5	300
6.3.4	2.3	1200
6.3.5	2.3	6000
6.4.2	4.3	1380
6.4.3	2.5	350

Tabelle 10, Vergleich Guete - Rechenzeit

Aus der Tabelle 10 geht hervor, dass das optimale Verhaeltnis von Rechenzeitaufwand zu erreichter Sprachqualitaet durch die Formantbestimmung mittels inverser Filterung der Sprachzeitfunktion nach Kap 6.4.3 erzielt wurde. Dieses Verfahren wurde deshalb auch fuer die weiteren Untersuchungen benutzt.

6.6 Amplitudenbestimmung

Ein weiterer, sehr wichtiger Parameter, der im Analyseteil des Formantvocoders bestimmt werden muss, ist neben der Pitchbestimmung und Formantbestimmung die Amplitudenbestimmung.
Der Amplitudenwert wird als Groesse ermittelt, die proportional dem Effektivwert ist und in ihrem Betrag den Wert 10000 nicht ueberschreiten kann. Der Wert 10000 entspricht gerade einer Maschineneinheit des Analogrechners in dem benutzten Hybridsystem.
Die Berechnung des Amplitudenfaktors erfolgt nach Gl.(97).

$$MA(t) = \sqrt{\frac{\int_{-\frac{I}{2}}^{\frac{I}{2}} w^2(\tau - t) \cdot f^2(\tau - t)\, d\tau}{\int_{-\frac{I}{2}}^{\frac{I}{2}} w^2(\tau)\, d\tau}} \qquad (97)$$

Darin bedeuten w(t) der Verlauf des verwendeten Zeitfensters und f(t) die analysierte Sprachzeitfunktion. Als Zeitfenster wird ein Hammingwindow mit einer Gesamtlaenge von 51.2 ms verwendet.

7. Manipulation der Steuerparameter

7.1 Glaettung der Parameterverlaeufe

Die Verlaeufe der Steuerparameter, wie sie unmittelbar aus der Analyse gewonnen werden, bestehen aus stark streuenden Einzelwerten. Das wird ganz besonders aus den Verlaeufen der Formantfrequenzen nach Abb.72 bis Abb.77 deutlich.

Die Streuung der Einzelwerte ist alleine auf die Unsicherheit der einzelnen Formantbestimmungsverfahren zurueckzufuehren und liegt nicht im Formantverlauf an sich begruendet. Synthetisiert man Sprache unmittelbar aus diesen Parameterverlaeufen, erhaelt man ein klimperndes Hintergrundsgeraeusch, dass teilweise genauso laut wie die eigentliche Sprache wird. Um diese zusaetzliche, in der Sprache nicht vorhandene Information wieder zu entfernen, muessen die Parameterverlaeufe geglaettet werden. Diese Glaettung darf aber wiederum nicht zu stark erfolgen, da, wie bereits in Kap 6.5 erwaehnt wurde, die Sprachinformation vorwiegend in den Aenderungen der Parameterverlaeufe enthalten ist.

Die Glaettung der Parameterverlaeufe wird vom Verfasser innerhalb bestimmter Zeitintervalle durchgefuehrt. Als die Zeitintervalle werden die gewaehlt, in denen konstant auf Stimmhaftigkeit bzw. Stimmlosigkeit entschieden wird. Als maximale Laenge fuer ein Intervall wurde die Laenge von 0.4s festgelegt.

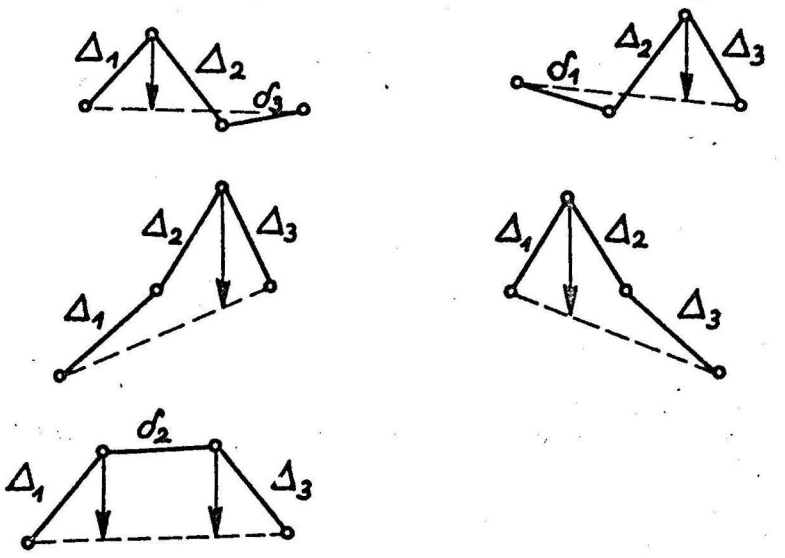
Da die Glaettung der Parameterverlaeufe fuer jedes Intervall getrennt und unabhaengig von den Nachbarintervallen durchgefuehrt wird, treten an den Intervallgrenzen groessere Parameterspruenge auf. Hierdurch soll eine zusaetzliche Haerte in den Klang der synthetisierten Sprache gebracht werden, die der deutschen Sprache im Gegensatz zur englischen ohnehin eigen ist.

Die Glaettung der Parameterverlaeufe innerhalb eines Intervalls wird in zwei Schritten durchgefuehrt: Zunaechst erfolgt eine nichtlineare Glaettung, mit der die sog. 'Ausreisser' in den Parameterverlaeufen erfasst werden sollen. Es wird zunaechst einmal der Mittelwert fuer die Differenzen benachbarter Werte ermittelt. Solche Punkte des Parameterverlaufes werden als moegliche Kandidaten fuer einen Ausreisser betrachtet, bei denen die Differenz zu einem ihrer Nachbarwerte das 1.5-fache des genannten Mittelwertes ueberschreitet.

Zur Beurteilung, ob ein Ausreisser vorliegt, werden immer drei Differenzen herangezogen. Die Abb.78 zeigt die Faelle, die als Ausreisser betrachtet werden und deutet durch gestrichelte Darstellung zugleich die Korrektur an. Das Zeichen δ deutet in der Abb.78 eine zulaessige Differenz und ein Δ eine unzulaessig grosse Differenz zwischen zwei

benachbarten Punkten an.

Im zweiten Schritt werden die Parameterverlaeufe innerhalb eines Intervalls durch Polynome so angenaehert, dass die mittlere quadratische Abweichung vom nichtlinear geglaetteten Parameterverlauf ein Minimum darstellt. Der Grad der Glaettung laesst sich durch die Ordnungszahl des Polynomens waehlen.



Die folgende Korrektur wird nur am Anfang bzw. Ende eines Intervalls durchgeführt

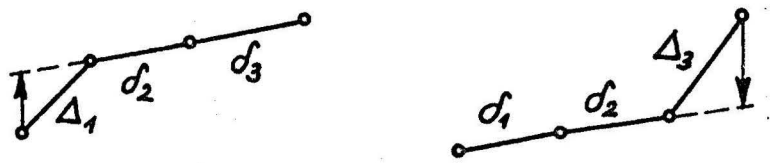


Abb.78, Nichtlineare Glaettung

Die Wahl einer festen Ordnungszahl haette zur Folge, dass kurze Intervalle kaum und lange Intervalle stark geglaettet wuerden. Die Ordnungszahl der Polynome wurde deshalb mit der Intervalllaenge gekoppelt.

7.2 Amplitudenregelung des Vocoders

Zur Sprachsynthese wird ein Formantsynthetisator nach Abb.48 verwendet. Die Steuerparameter werden getrennt nach Pitchbestimmung, Formantbestimmung und Amplitudenbestimmung so berechnet, wie es in Kap 6 beschrieben wurde. Nach 7.1 werden die Parameterverlaeufe bis auf die Stimmhaft-Stimmlos-Entscheidung LVU geglaettet und sollen jetzt unmittelbar zur Steuerung des Formantvocoders verwendet werden. Das Ergebnis einer derartigen Synthese zeigt, dass nur in den wenigsten Faellen auf diese Weise ein vernuenftiges Ergebnis erzielt werden kann. In den meisten Faellen treten Dynamikschwankungen des Ausgangssignals auf, die teilweise groesser als 80dB sind. Von der Sprache ist dann am Ausgang nichts mehr zu hoeren, da der Ausgangsverstaerker eingangsseitig auf die maximal auftretenden Amplituden eingeregelt werden muss.

Die Ursache fuer die Dynamikschwankungen kann man anschaulich sehr einfach deuten: Die Formantbestimmung und die Amplitudenbestimmung sind unabhaengig voneinander durchgefuehrt worden, obwohl sie eigentlich recht eng miteinander verknuepft sind. Die Ausgangsamplitude eines zu Schwingungen angeregten Formantnetzwerkes ist dann am kleinsten, wenn die Frequenzwerte der Formanten moeglichst weit voneinander ent-

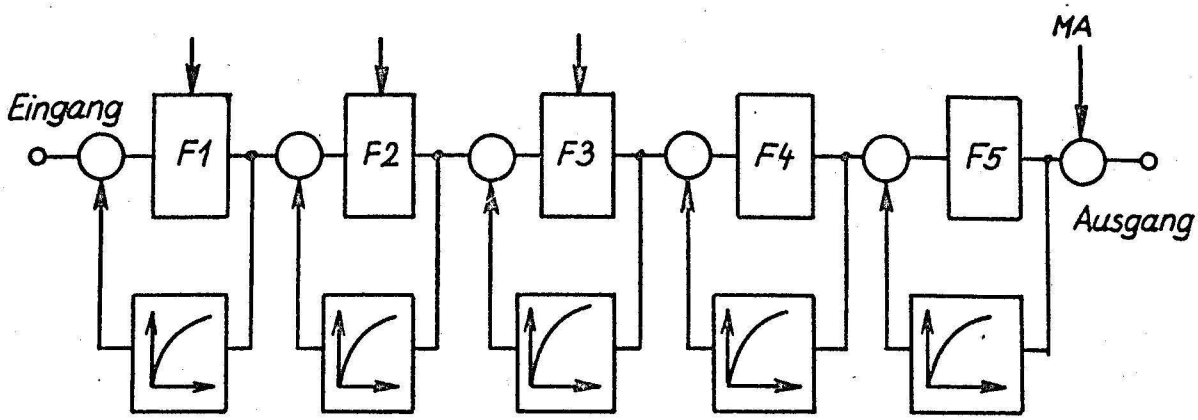


Abb.79, Frequenzunabhaengige Amplitudensteuerung des Formantfilters

fernt liegen. Liegen die Frequenzwerte der Formanten dagegen dicht beieinander, kommt es zu einer starken Resonanzueber-

hoehung der Ausgangsamplitude. Es gibt nun zwei Moeglichkeiten, wie man das vorliegende Problem loesen kann.

1. Im Formantfilter nach Abb.48 erhaelt jedes Formantglied eine Dynamikregelung, die am Ausgang jedes Formanten ein Signal erzeugt, dessen Amplitude unter Beruecksichtigung einer gewissen Traegheit der Regelschaltung unabhaengig von der Frequenz ist. Der gewuenschte Amplitudenverlauf kann dann, wie aus Abb.79 ersichtlich ist, am Ausgang als Parameter MA der synthetisierten Sprache eingepraegt werden.
   Diese Loesung erfordert bei einer hardwaremaessigen Realisierung des Synthetisators einen zu grossen zusaetzlichen Aufwand.

2. Die zweite Moeglichkeit ist in Abb.80 angedeutet. Die Amplitude am Ausgang des fuenften Formanten wird ermittelt und damit der Eingang des ersten Formanten ueber eine Regelschaltung ausgesteuert. Bei dieser Realisierung wird zwar nur eine Regelschaltung benoetigt, jedoch treten zwischen den Formantgliedern recht grosse und ausserordentlich kleine Amplituden auf. Diese Art der Regelung ist nur dann vertretbar, wenn die Formantglieder in Gleitkomma- oder zumindest in Block-Floatingpoint-Arithmetik ausgefuehrt sind.

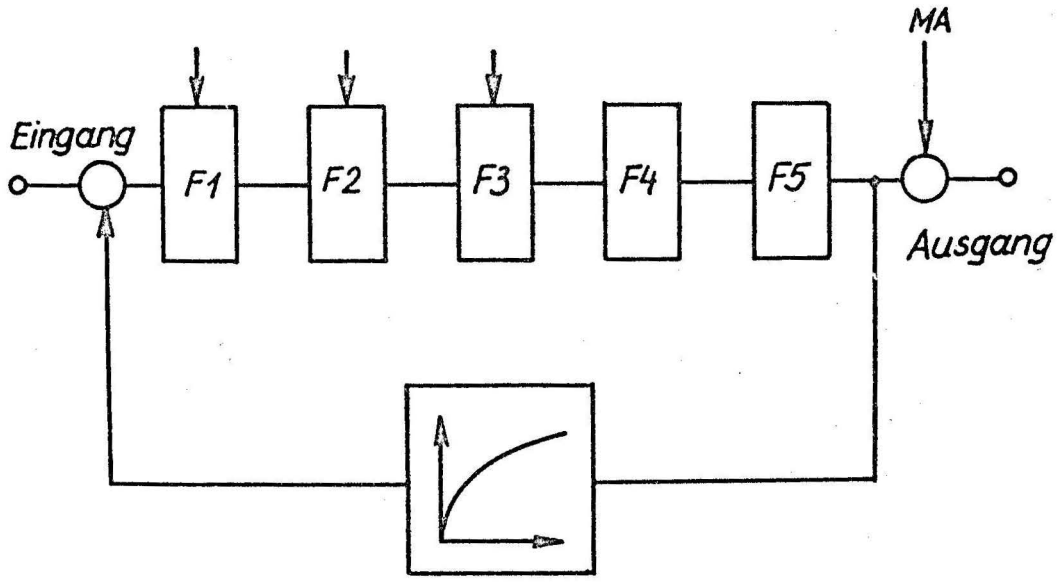


Abb.80, Frequenzunabhaengige Amplitudensteuerung des Formantfilters

Der Verfasser hat die zweite Moeglichkeit mit einer Abwandlung verwendet: Aufgrund der Linearitaet des Formantfilters

braucht die Dynamikregelung nicht am Eingang des ersten Formanten zu erfolgen, sondern kann mit der Aussteuerungsregelung von MA zusammengelegt werden. Die Regelung braucht ausserdem nicht waehrend der Synthese selbst stattzufinden, sondern kann bereits bei der Parameteraufbereitung beruecksichtigt werden. Das wird dadurch erreicht, dass das Formantnetzwerk simuliert und damit der Amplitudenverlauf am Ausgang des fuenften Formanten als MF5 ermittelt wird. Die neue Amplitudengroesse MA berechnet sich zu

$$MA = K \cdot ln \frac{MA^*}{MF5} \qquad (98)$$

In Gl.(98) stellt MA^* den nach 6.6 ermittelten Effektivwert dar und MF5 den ueber 51.2 ms gemittelten Effektivwert am Ausgang des fuenften Formanten, der bei der Zweitsimulation ermittelt wurde. K stellt eine Konstante dar.

Da in Gl.(98) der Logarithmus verwendet wurde, muss bei einer Synthese nach Abb.48 eine exponentielle Amplitudenregelung bei Verwendung des Parameters MA statt MA^* erfolgen. Das kommt einer Hardwarerealisierung in Digitaltechnik sehr entgegen, da statt einer Multiplikation mit dem Amplitudenfaktor lediglich der Inhalt des Ausgangsregisters geschiftet werden muss.

7.3 Codierung

Die Codierung der Parameterverlaeufe erfolgt aus Tabellen. Fuer jeden Parameter wird eine Codierungstabelle mit 2**NBIT Werten berechnet, wobei NBIT die zur Codierung des Parameters gewuenschte Bitzahl angibt.

Bei der Berechnung der Tabelle ging der Verfasser von der Annahme aus, dass die Parameterverlaeufe sich mehr oder weniger um einen Mittelwert konzentrieren und dass Parameterwerte um so seltener auftreten, je weiter sie von diesem Mittelwert entfernt liegen. Ausserdem kann man einen groessten Wert XMAX und einen kleinsten Wert XMIN angeben, die auf keinen Fall ueberschritten bzw. unterschritten werden. Aus einer grossen Anzahl von Sprachbeispielen wurde fuer die einzelnen Parameter der Mittelwert XQUER und die Streuung SIGMA ermittelt. Daneben wurden die oberen und unteren Grenzen festgestellt. Die Werte sind, zum Teil abgerundet, in der Tabelle 11 dargestellt.

		XQUER	SIGMA	XMIN	XMAX
Pitchquefrenz	10^{-4} sek	85	40	43	150
Amplitude	dB	42	40	0	68
1. Formant	Hz	500	1000	200	1500
2. Formant	Hz	1765	1000	652	2812
3. Formant	Hz	2500	500	2000	3100

Tabelle 11, Statistische Angaben ueber die Steuerparameter

Unter Vorgabe von NBIT wurde fuer alle Parameter getrennt je eine Tabelle mit 2**NBIT Werten erstellt, in der die Abstaende benachbarter Parameterwerte umgekehrt proportional zu ihrer Haeufigkeit auftreten. Als maximale Codierungszahl wurde NBIT=9 vorgesehen.

Codiert man die Parameterverlaeufe lediglich dadurch, dass man die einzelnen Parameterwerte einer Tabelle entnimmt, erhaelt man immer noch eine zu grosse Redundanz. Man kann beispielsweise den Formantverlaeufen nach Abb.72 bis Abb.77 entnehmen, dass diese kontinuierlich verlaufen. Der kontinuierliche Verlauf innerhalb eines stimmhaften bzw. stimmlosen Bereiches wird dadurch gefoerdert, dass die Parameterverlaeufe innerhalb dieses Bereiches durch Polynomenverlaeufe approximiert worden sind. Das bedeutet, dass man nur den ersten Wert innerhalb eines stimmhaften oder stimmlosen Bereiches voll mit NBIT bit zu codieren braucht und dann immer nur noch die Differenz zum vorhergehenden Wert durch die geringere Codierung mit NDBIT bit uebertragen zu

braucht.
Eine weitere Redundanzminderung kann dadurch erreicht werden, dass in stimmlosen Bereichen die Pitchfrequenz nicht uebertragen zu werden und waehrend der Zwischenraeume, d.h. in Bereichen mit LVU=2 gar keine Parameter uebertragen zu werden brauchen. Die Gesamtcodierung setzt sich damit folgendermassen zusammen:

1. Fuer einen stimmhaften Bereich

Stimmhaft-Stimmlosigkeit	LVU $\hat{=}$ 2 bit
Laenge des Bereiches	NL $\hat{=}$ 6 bit
Volle Codierung des ersten Wertes fuer alle 5 in Tabelle 11 aufgefuehrten Parameter	$\sum_{i=1}^{5} NBIT_i$
Codierung der folgenden Werte	$(NL-1)\cdot\sum_{i=1}^{5} NDBIT_i$

2. Fuer einen stimmlosen Bereich

Stimmhaft-Stimmlosigkeit	LVU $\hat{=}$ 2 bit
Laenge des Bereiches	NL $\hat{=}$ 6 bit
Volle Codierung des ersten Wertes fuer alle in Tabelle 11 aufgefuehrten Parameter bis auf den ersten.	$\sum_{i=2}^{5} NBIT_i$
Codierung der folgenden Werte	$(NL-1)\cdot\sum_{i=2}^{5} NDBIT_i$

3. Fuer einen Zwischenraum

Stimmhaft-Stimmlosigkeit	LVU $\hat{=}$ 2 bit
Laenge des Bereiches	NL $\hat{=}$ 8 bit

Der Quantisierungsfaktor NBIT und der Faktor zur Codierung der Aenderung des Parameterverlaufes, NDBIT, muessen fuer jeden der 5 Parameter getrennt ermittelt werden, da die Parameter mit unterschiedlichem Gewicht zur Verstaendlichkeit der Sprache beitragen. In dem Zusammenhang wurden vom Verfasser umfangreiche Untersuchungen angestellt, in denen Sprachbeispiele mit verschiedensten Werten fuer NBIT und NDBIT synthetisiert wurden. Ausserdem wurde untersucht, welchen Einfluss eine Parameteruebergabe alle 20 ms im Gegensatz zu einer Parameteruebergabe alle 10 ms an den Synthetisator auf die Sprachqualitaet hat. Die Ergebnisse der Untersuchung sind die folgenden:
Die Bedeutung der Parameter ist ihrer Reihenfolge nach:

1.) Amplitudenkontrolle
2.) 1. Formantfrequenz
3.) 2. Formantfrequenz
4.) 3. Formantfrequenz
5.) Pitchfrequenz

Die Amplitudenkontrolle hat deshalb ein so grosses Gewicht, da sie nicht lediglich den Verlauf des Effektivwertes beinhaltet, sondern auch noch die in Kap 7.2 erwaehnten Regelaufgaben wahrnehmen muss. Es hat sich ergeben, dass zur Quantisierung der Amplitudenkontrolle etwa 1 bis 2 bit mehr notwendig sind, als zur Quantisierung des ersten Formanten. Aufgrund der Regelwirkung der Amplitudenkontrolle muss mit schnellen Aenderungen dieses Parameters gerechnet werden, so dass man zweckmaessigerweise NDBIT=NBIT waehlt.

Der erste und zweite Formant spielen die wesentliche Rolle bei der Formung von Lauten innerhalb einer Lautkategorie. Da es beim zweiten Formanten vor allem auf die richtige Tendenz des Frequenzverlaufes, beim ersten Formanten dagegen etwas mehr auch auf den richtigen Frequenzwert ankommt, kann man den zweiten Formanten um 1 bit niedriger als den ersten Formanten quantisieren.

Der dritte Formant, der in seiner Bedeutung noch unter der des zweiten Formanten steht, kann noch um ein weiteres Bit geringer quantisiert werden, ohne zu einer einseitigen Verschlechterung der Sprachqualitaet zu fuehren.

Die Formantfrequenzverlaeufe sind i.a. kontinuierliche Verlaeufe, so dass NDBIT<NBIT gewaehlt werden kann. In dem Kap 6.5 wurde aber bereits erwaehnt, dass nach Ansicht des Verfassers die Sprachinformation gerade in den Aenderungen der Parameterverlaeufe steckt und es wurde an der Formantbestimmung nach Kap 6.4.2 kritisiert, dass ein zu 'flacher' Parameterverlauf zu einer fast unverstaendlichen Sprache fuehrt. Die Aenderung der Parameterverlaeufe, die durch NDBIT dargestellt wird, darf deshalb nicht zu klein gemacht werden. Synthesebeispiele haben gezeigt, dass man fuer die drei Formantfrequenzen vorteilhaft NDBIT=NBIT-2 waehlt, wenn die Parameteruebergaben an den Synthetisator alle 10 ms und NDBIT=NBIT-1, wenn die Parameteruebergabe alle 20 ms erfolgt.
Der Verfasser waehlte den letzten Fall, weil durch eine Parameteruebergabe aller 20 ms die Anzahl der Bits zur Gesamtcodierung erheblich verringert werden konnte.

Am langsamsten aendert sich von allen Parametern die Pitchfrequenz. Man kann in jedem Fall NDBIT=NBIT-2 waehlen. Bei der in Tabelle 12 angegebenen Codierung konnte nur eine sehr geringe Qualitaetseinbusse gegenueber der aus uncodierten Parameterverlaeufen synthetisierten Sprache festgestellt werden. Die zur Uebertragung benoetigte Bitrate betraegt ca 1000 - 1200 bit/sek.
Es muss an dieser Stelle darauf hingewiesen werden, dass saemtliche Untersuchungen mit deutscher Sprache durchgefuehrt wurden. Der Verfasser ist der Ansicht, dass zur Synthese englischer Sprache, die in ihrem Klang weitaus weniger

	NBIT	NDBIT
Pitchfrequenz	5	3
Amplitude	8	8
1. Formant	6	5
2. Formant	5	4
3. Formant	4	3

Tabelle 12, Codierungsbeispiel

hart als die deutsche ist, bei vergleichbarer Sprachqualitaet weniger Bit zur Codierung benoetigt wuerden. Diese Behauptung muss jedoch erst durch Untersuchungen gestuetzt werden.

8. Hardwareausfuehrung eines Formantsynthetisators

Fuer den simulierten Formantvocoder nach Abb.48 soll eine hardwaremaessige Realisierung gefunden werden, die von einem Digitalrechner gesteuert werden kann. Es gibt dafuer prinzipiell zwei verschiedene Moeglichkeiten:

1. Analoge Ausfuehrung
2. Digitale Ausfuehrung

Im folgenden soll fuer die beiden Moeglichkeiten je ein Realisierungsvorschlag skizziert werden.

8.1 Analoge Ausfuehrung

Der Formantsynthetisator besteht aus dem Anregungsgenerator und dem Formantfilter. Der Anregungsgenerator enthaelt umschaltbar einen Rauschgenerator und einen Pulsgenerator, der in seiner Pulsfolgefrequenz vom Digitalrechner regelbar sein muss. Als Rauschgenerator kann das Rauschen einer Zenerdiode verwendet werden, aus dessen Rauschspektrum ein ca 4.5 kHz

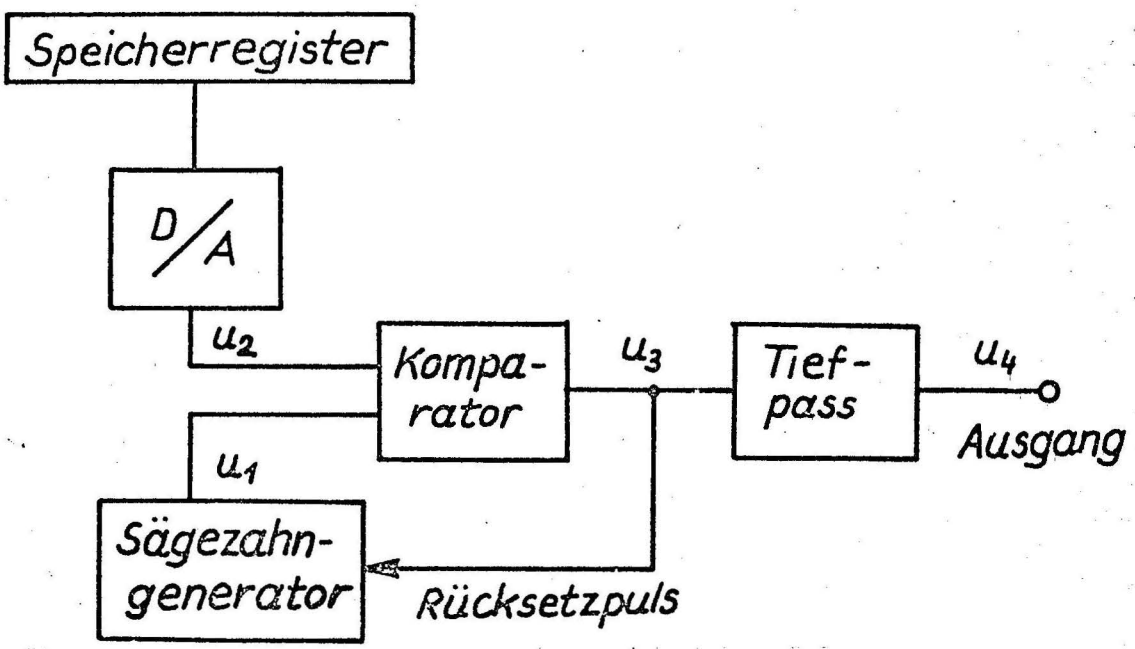


Abb.81, Blockschaltbild eines digital steuerbaren Pulsgenerators

breites Frequenzband durch Modulation und nachfolgende Tief-

passfilterung herausgegriffen wird. Ein derartiger Rauschgenerator, der ein weisses Rauschen mit einer Gaussverteilung der Amplituden liefert, wurde vom Verfasser /44/ bereits beschrieben.
Den prinzipiellen Aufbau eines digital steuerbaren Pulsgenerators zeigt die Abb.81. Die zugehoerigen Spannungsverlaeufe sind der Abb.82 zu entnehmen. Ein Saegezahngenerator erzeugt eine zeitproportionale Spannung $u_1(t)$. Die gewuenschte Periodenlaenge des Pulsgenerators wird ueber einen D/A-Umsetzer in die ihr proportionale Spannung $u_2(t)$ verwandelt. Ein Komparator vergleicht $u_1(t)$ und $u_2(t)$. Wenn $u_1(t)$ groesser als $u_2(t)$ wird, liefert der Komparator einen Ausgangspuls $u_3(t)$, der zugleich den Saegezahngenerator zuruecksetzt. Der Abstand T_p benachbarter Pulse ist proportional der Vergleichsspannung $u_2(t)$ und damit proportional dem Inhalt des Speicherregisters. Durch einen Tiefpass kann der Pulsverlauf am Ausgang des Komparators in einen dreieckfoermigen Verlauf entsprechend $u_4(t)$ in Abb.82 umgewandelt werden.

Die Umschaltung des Rauschgenerators auf den Pulsgeneratorzweig bzw. die Abschaltung des Anregungsgenerators vom Formantfilter kann vom Rechner ueber Controllines erfolgen, die schnelle Relais schalten. Es reichen dabei Schaltzeiten von 10 ms fuer das Relais aus.

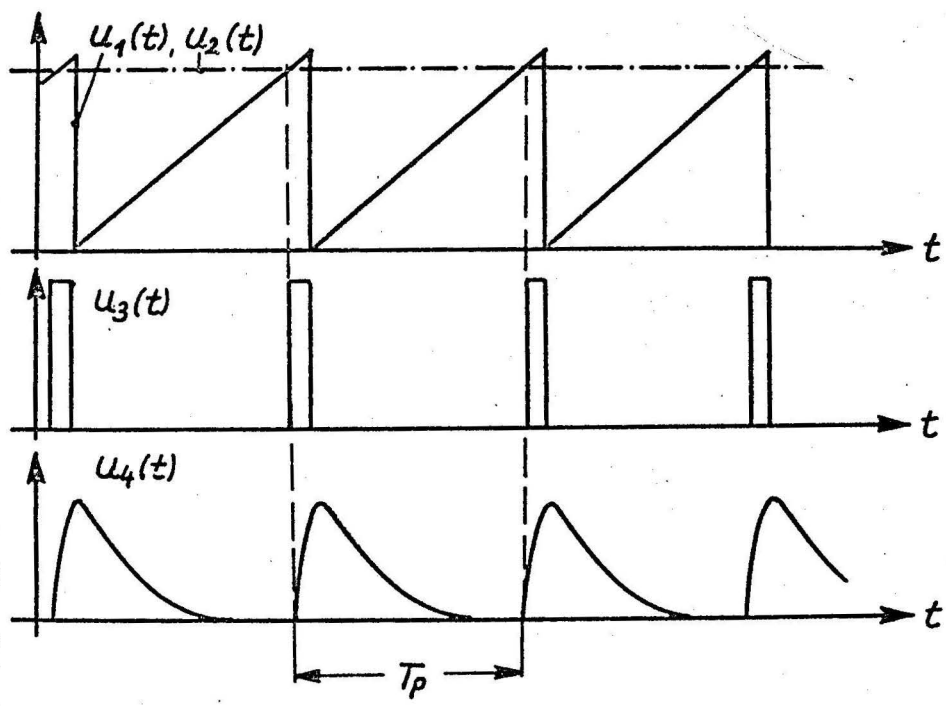


Abb.82, Spannungsverlaeufe im Pulsgenerator

Fuer den Aufbau eines Formantgliedes wurde vom Verfasser ein Aufbau nach Abb.85 ausprobiert. Die Anregung dazu stammt aus einer Veroeffentlichung von BRONZITE /45/ unter dem Titel 'Audio-spectrum-analyzer'.

Ein RC-Filter mit Bandpasscharakter nach Abb.83 hat die Resonanzfrequenz

$$f_r = \sqrt{n}/2\pi RC$$

und bei der Resonanz die Verstaerkung

$$g_r = 2(2+n)$$

wobei n das Verhaeltnis der Kapazitaeten ist. Durch Veraen-

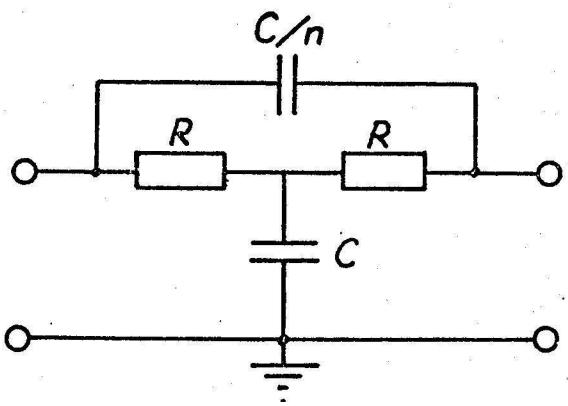


Abb.83, RC-Filter mit Bandpasscharakter

derung der beiden Widerstaende R laesst sich die Resonanzfrequenz variieren, waehrend die Verstaerkung g_r davon unabhaengig bleibt. Die Veraenderung der Widerstandswerte wird durch Modulation ueber FET-Schalter entsprechend Abb.84 erreicht. Das Modulationssignal besteht aus Pulsen der Pulsfolgefrequenz im Bereich von 50 - 500 kHz. Das Tastverhaelt-

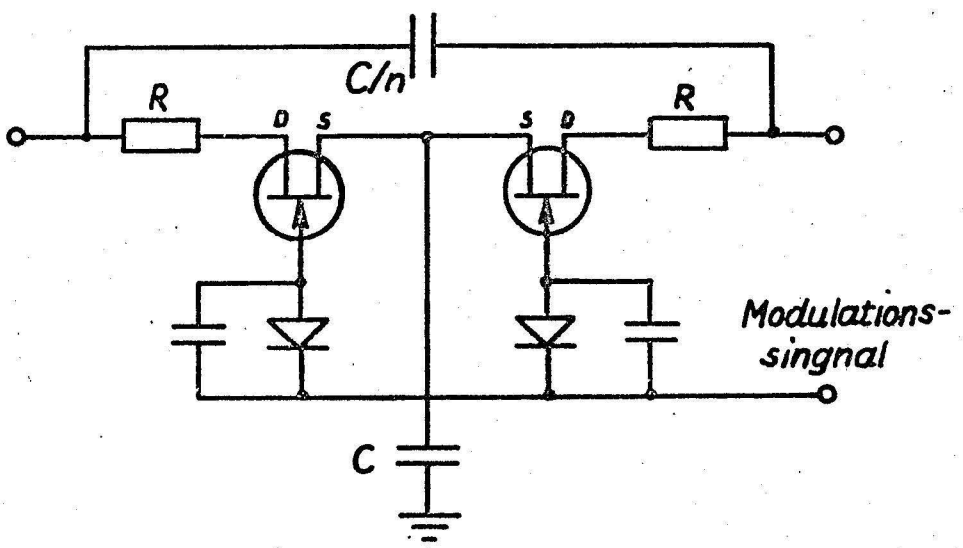


Abb.84, Aenderung der Widerstandswerte durch FET-Schalter

nis der Pulse kann in linearer Abhaengigkeit von einer Steuerspannung im Verhaeltnis 1 : 20 variiert werden. Das bedeutet, dass die Resonanzfrequenz des Filters im Verhaeltnis 1 : 20 geaendert werden kann. Die Schaltung wurde so dimensioniert, dass ein Frequenzbereich von 250 Hz bis 5 kHz ueberstrichen werden kann. Das RC-Filter wurde nach BRONZITE mit einem Operationsverstaerker nach Abb.85 zusammengeschaltet. Das HF-Filter am Ausgang der Formantschaltung dient da-

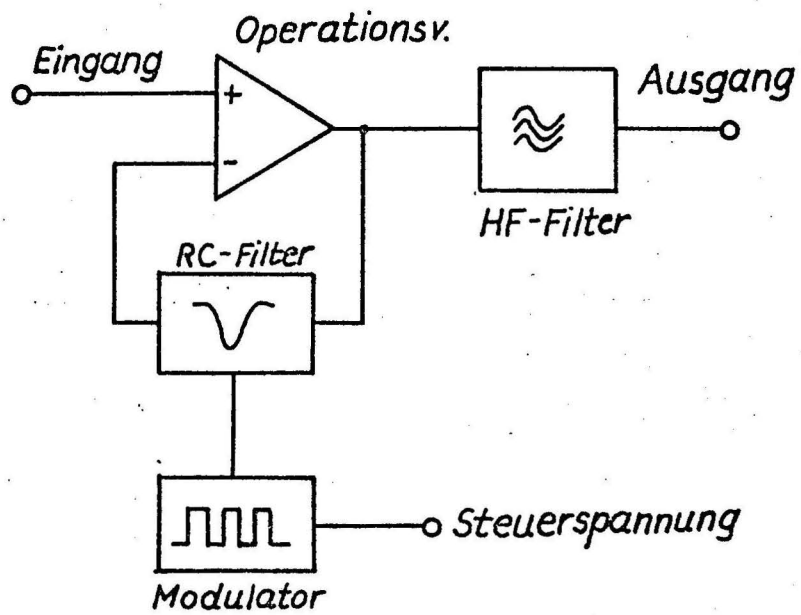


Abb.85, Blockschaltbild eines Formantgliedes

zu, die hochfrequenten Anteile, die vom Modulator herruehren, aus dem Ausgangssignal fernzuhalten. Die Einstellung eines Daempfungswertes unabhaengig von der Frequenz ist bei dieser Schaltung nicht moeglich. Da die Bandbreitewerte der Formanten ohnehin nur einen geringen Einfluss auf die Verstaendlichkeit der Sprache haben, ist dieser Nachteil hier ohne Bedeutung.

Auf eine andere Moeglichkeit, ein Formantnetzwerk durch eine Analogschaltung zu realisieren, wurde bereits in Kap 5.6 hingewiesen.

Dem Vorteil einer relativ einfachen Realisierungsmoeglichkeit, sowohl der einzelnen Formantglieder, als auch des Generators steht als grosser Nachteil die umstaendliche und aufwendige Steuerung eines analogen Formantsynthetisators gegenueber.

Beim Aufbau des Formantfilters nach Abb.48 treten Schwierig-

keiten bei der Aussteuerung der Formantglieder auf, die schon in Kap 5.6 bei der Synthese auf dem Hybridrechner behandelt wurden. Bei einer seriellen Ausfuehrung des Formantfilters werden zur Aussteuerungsregelung 5 multiplizierende Digital-Analog-Umsetzer und 4 Analog-Digital-Umsetzer zur Erfassung der Amplituden an den Ausgaengen der ersten vier Formantnetzwerke benoetigt, wenn man eine Regelung vornehmen will, die der im Kap 5.6 beschriebenen entspricht. Ausserdem muesste entweder eine komplizierte digitale Steuerschaltung fuer die multiplizierenden DAU's aufgebaut werden, oder die Steuerung vom Digitalrechner uebernommen werden.

Eine Alternative zur Aussteuerungsregelung der Formantglieder waere die, fuer jeden Formanten eine eigene Dynamikregelung vorzusehen und lediglich am Ausgang ueber einen multiplizierenden DAU der Sprache den richtigen Amplitudenverlauf einzupraegen. Der Verfasser hat aber noch nicht untersuchen koennen, mit welchen Schwierigkeiten eine Reihenschaltung von fuenf voneinander unabhaengigen, in der Dynamik gesteuerten Formantgliedern versehen ist, und es ist sehr fraglich, ob auf diese Weise ueberhaupt ein stabiler Aufbau des Formantfilters moeglich ist.

Eine sehr viel einfachere Realisierungsmoeglichkeit fuer ein analoges Formantfilter bietet eine Parallelschaltung der fuenf Formantnetzwerke. Das bedeutet wiederum, dass jeder Zweig nicht nur eine eigene Amplitudenregelung benoetigt, sondern auch, dass wesentlich mehr Parameter fuer die Sprachsynthese vom Rechner gespeichert werden muessen, denn, wie schon in Kap 5.1 erwaehnt wurde, muessen diese Koeffizenten im Zusammenhang mit der Sprachanalyse berechnet werden.

8.2 Digitale Ausfuehrung

Beim Aufbau eines digitalen Formantsynthetisators sind die Verhaeltnisse ganz anders als bei der analogen Ausfuehrung. Die Steuerung des digitalen Formantsynthetisators ist verhaeltnismaessig einfach, da alle Signale bereits in digitalisierter Form vorliegen. Die Schwierigkeiten, die sich hier nur mit hohem Aufwand realisieren lassen, ist der Aufbau des Formantelementes selbst. Eine Formantschaltung laesst sich fuer den kontinuierlichen Fall durch eine einfache Differentialgleichung, fuer den diskreten Fall durch eine einfache Differenzgleichung darstellen. Es ist aber sehr viel einfacher, Rechenoperationen wie Integration und Addition mit Methoden der Analogtechnik auszufuehren als z.B. eine Addition und eine Multiplikation in Digitaltechnik. Der Verfasser sieht fuer die Zukunft trotzdem einen grossen Vorteil in der Entwicklung eines rein digitalen Formantsynthetisators, da zum einen mit der Entwicklung groesserer Read-Only-Memories saemtliche Rechenoperationen zwischen zwei Operanden, also auch die Multiplikation, mit Hilfe einer Tabelle einschrittig ausgefuehrt werden koennen. Zum anderen wird es durch Einfuehrung immer billigerer und schnellerer Kleinstrechner sicherlich bald moeglich sein, solche Problemstellungen, wie den Aufbau eines Formantsynthetisators durch einen festprogrammierten Minicomputer in Realzeit zu loesen. In beiden Faellen wird ein rein digita-

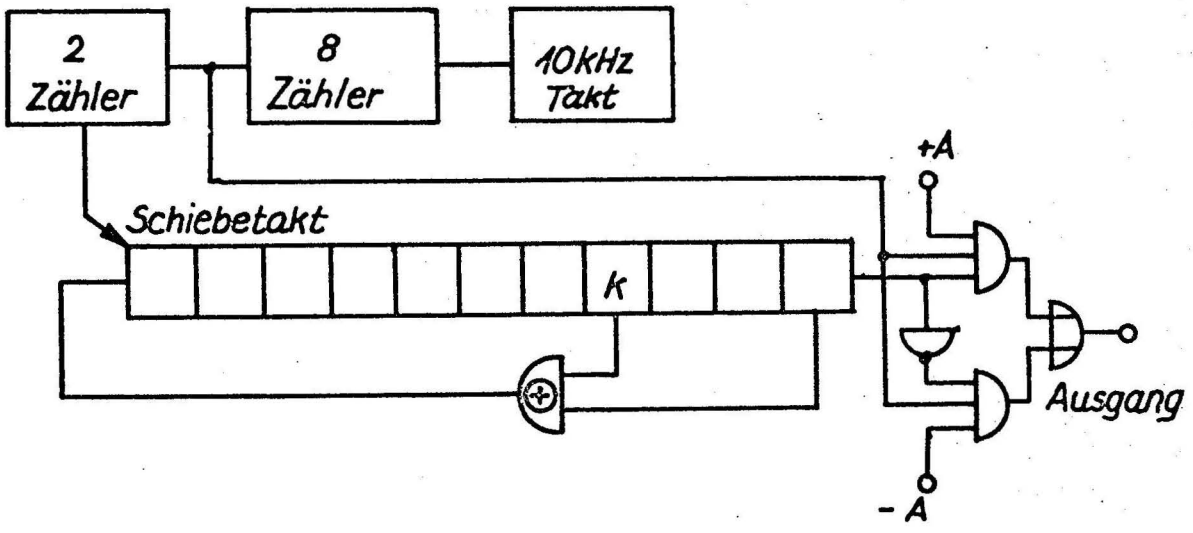


Abb.85, Schieberegister-Rauschgenerator

les Konzept benoetigt, das hier im folgenden angedeutet werden soll.

Der Rauschgenerator besteht aus einem Schieberegisterrauschgenerator nach Abb.86. Der Ausgang des Schieberegisters und der k-te Abgriff werden ueber eine Disvalenz verknuepft und auf den Eingang zurueckgegeben. Schreibt man in diese Schaltung einen bestimmten, von Null verschiedenen Anfangszustand ein, so liefert jede Stelle des Schieberegisters nach Anlegen des Schiebetaktes eine Ausgangsgroesse, die aus einer Folge unabhaengiger binaerer Zufallszahlen besteht. Die Folge wiederholt sich nach 2**N-1 Takten. Es gibt nur bestimmte Schieberegisterlaengen N, bei denen mit einer einzigen Rueckfuehrung an der Stelle k<N eine Folge der maximal moeglichen Laenge moeglich ist. Die Tabelle 13 gibt eine Uebersicht ueber verschiedene Registerlaengen N, bei denen lediglich ein Abgriff .k<N benoetigt wird, die Lage des Abgriffspunktes k und die Periodenlaenge.

Registerlaenge	Abgriff	Periodenlaenge
11	2	2 047
15	1	32 767
17	3	131 071
18	7	262 143
20	3	1 048 575
21	2	2 097 151
22	1	4 194 303
23	5	8 388 607
25	3	33 554 431
28	3	268 435 455
31	3	2 147 483 647
33	13	8 589 934 591

Tabelle 13, Registerlaenge (bit), Abgriff und Periodenlaenge (Taktschritte) bei einem Schieberegisterrauschgenerator nach Abb.86

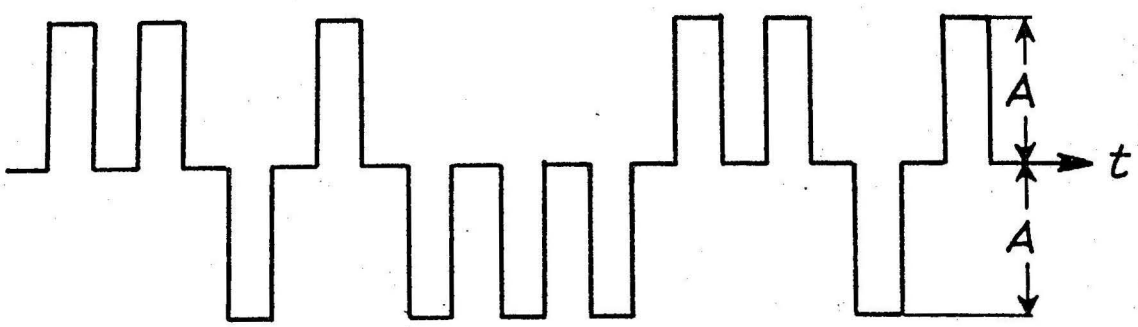


Abb.87, Zeitverlauf am Ausgang des Rauschgenerrators

Um eine guenstige Aussteuerung des nachfolgenden Formantnetzwerkes zu ermoeglichen, wird ein Rauschgenerator verwendet, der eine Folge von Pulsen der konstanten Amplitude A erzeugt, bei denen die Zufallsinformation im Vorzeichen der Pulse liegt. Der Schiebetakt wird durch Herabsetzung des 10 kHz-Taktes mit einem 16-Zaehler gewonnen. Die Pulsfolgefrequenz betraegt damit 625 Hz. Je nach dem Inhalt des letzten Bits des Schieberegisters wird die Amplitude +A oder die Amplitude -A auf den Ausgang des Rauschgenerators durchgeschaltet.
Der Ausgang ist mit 8 bit quantisiert.

Der prinzipielle Aufbau des Pulsgenerators ist in Abb.88 dargestellt. Die Laenge der Pitchperiode wird vom Digitalrechner in ein 8-bit-Register uebernommen. Ein 8-bit-Zaehler wird vom 10 kHz-Takt gesteuert. Wenn die beiden Register in

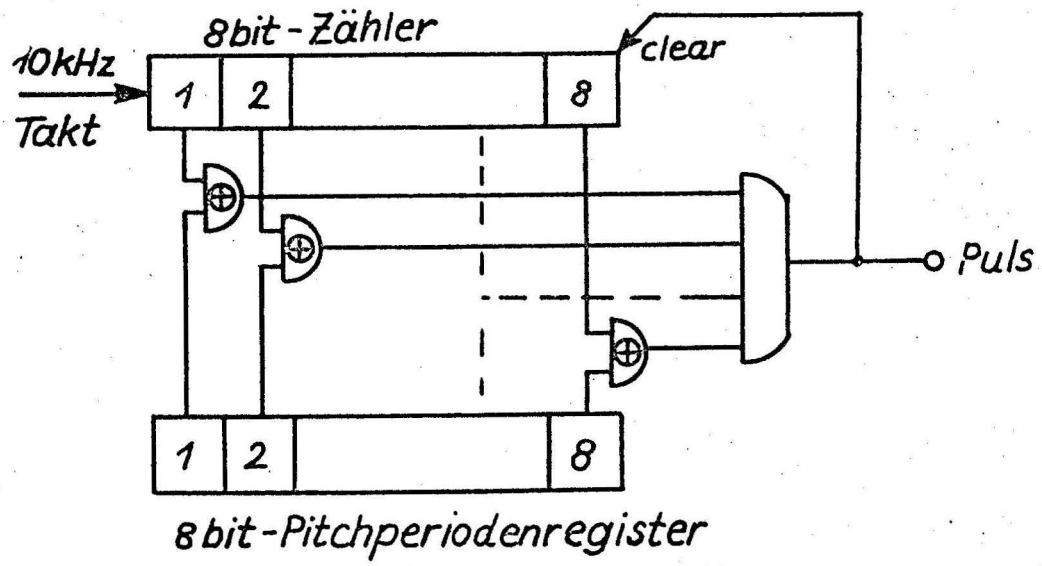


Abb.88, Pulsgenerator

ihrem Inhalt uebereinstimmen, wird durch eine Vergleichsschaltung der 8-bit-Zaehler auf den Anfang zurueckgesetzt und ein Puls abgegeben. Der Puls wird entweder direkt mit 8 bit quantisiert auf den Ausgang gegeben oder startet eine Pulsformschaltung. Die Pulsformschaltung soll aus dem Puls einen dreieckfoermigen Verlauf mit einer steilen Anstiegsflanke und einem flachen Abfall erzeugen. Das kann im einfachsten Fall durch einen Vor- Rueckwaertszaehler erfolgen, der mit dem 10 kHz-Takt beispielsweise 10 Schritte hochzaehlt und mit einem 5 kHz-Takt wieder bis auf Null zurueckzaehlt.

Der Frequenzbereich des Pulsgenerators ist nach unten durch die Laenge des Pitchperiodenregisters auf die Frequenz 10000/256=39 Hz und nach oben durch die Laenge des Dreieckspulses auf 10000/30=333 Hz begrenzt.

Mit $r=\exp(-\sigma_p \cdot T)$ und $b=\omega_p$ ergibt sich aus Gl.(48) die Uebertragungsfunktion eines Formanten zu:

$$H(z) = \frac{1-2r\cos bT}{1-2r\cdot\cos bT\, z^{-1}+r^2 z^{-2}} \qquad (99)$$

und aus Gl.(49) die zugehoerige Differenzengleichung zu:

$$y(nT) = 2r\cos bT\, y(nT-T) - r^2 y(nT-2T) + (1-2r\cos bT + r^2)\, x(nT) \qquad (100)$$

Die Gl.100 enthaelt drei Multiplikationen und zwei doppelte Additionen. Gl.(100) laesst sich folgendermassen in Gl.(101) umschreiben:

$$y(nT) = r^2[x(nT) - y(nT-2T)] + 2r\cos bT\,[y(nT-T) - x(nT)] + x(nT) \qquad (101)$$

Die Faktoren r^2 und $2\cdot r\cdot\cos bT$ werden der Formantschaltung vom Digitalrechner als Koeffizienten uebergeben.

Die Gl.(101) hat gegenueber der Gl.(100) den grossen Vorteil, dass sie nur zwei Multiplikationen, dafuer aber zwei einfache und eine doppelte Addition enthaelt. Das Blockschaltbild eines Formantgliedes nach Gl.(101) zeigt die Abb.89. Die einfache Struktur des Formantfilters nach Abb.48

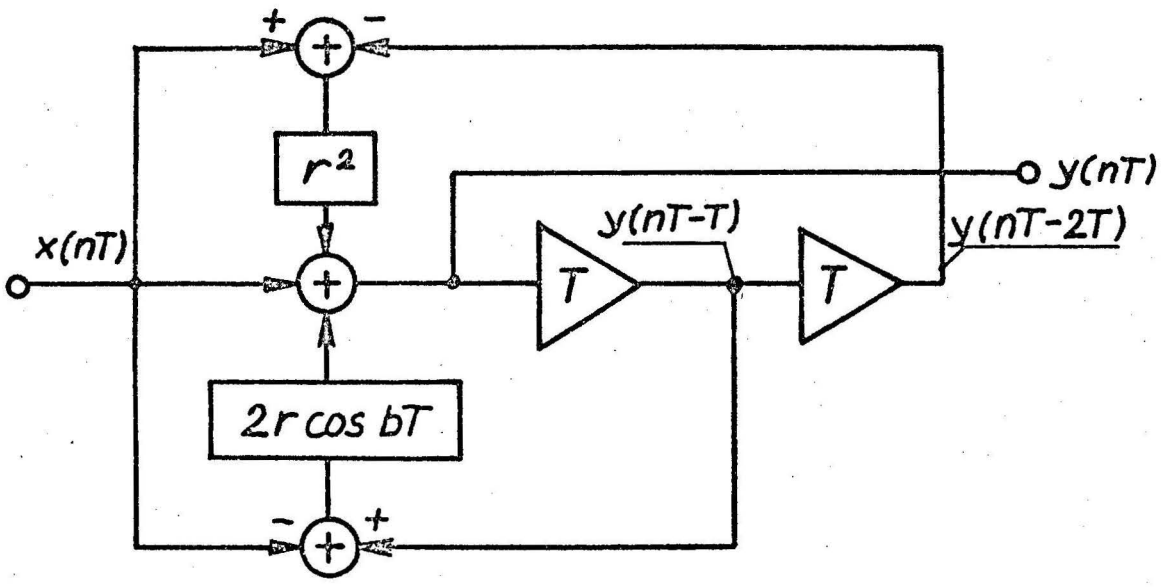


Abb.89, Formantschaltung mit zwei Multiplizierern

gestattet es, dass das Formantfilter nach Abb.7 nur einmal aufgebaut werden braucht und dann fuenffach gemultiplext werden kann.

Durch die Quantisierung des Signals, insbesondere durch die verwendete abgekuerzte Multiplikation tritt ein Quanti-

sierungsrauschen auf. Dieses Rauschen wurde theoretisch und experimentell von GOLD und RABINER /46/ untersucht. Es wurden durch Experimente mit einem Formantsynthetisator, der aus fuenf in Reihe geschalteten Formantgliedern bestand, die Registerlaengen fuer einzelne Vokale ermittelt, bei denen das Rauschen gerade an der Hoerschwelle lag. Fuer Formantnetzwerke entsprechend Gl.(100) wurden Registerlaengen bis zu 17 bit als notwendig ermittelt. Da als Rauschquelle vor allem die Multiplizierer in Frage kommen, ist der Verfasser der Meinung, dass bei einer Formantschaltung mit nur zwei Multiplizierern eine Registerlaenge von 16 bit ausreichen muesste.

Die Taktfrequenz fuer den Formantsynthetisator betraegt 10 kHz, bzw. die Zeitspanne zwischen zwei Ausgabewerten betraegt 100 µs. Die Rechenzeit fuer einen einzelnen Formanten darf also nur 20 µs betragen.

In Abb.90 ist die Formantschaltung nach Abb.89 noch einmal in etwas anderer Weise dargestellt. Die beiden Addierer ADD1 und ADD2 arbeiten parallel. Nach 1 µs koennen die

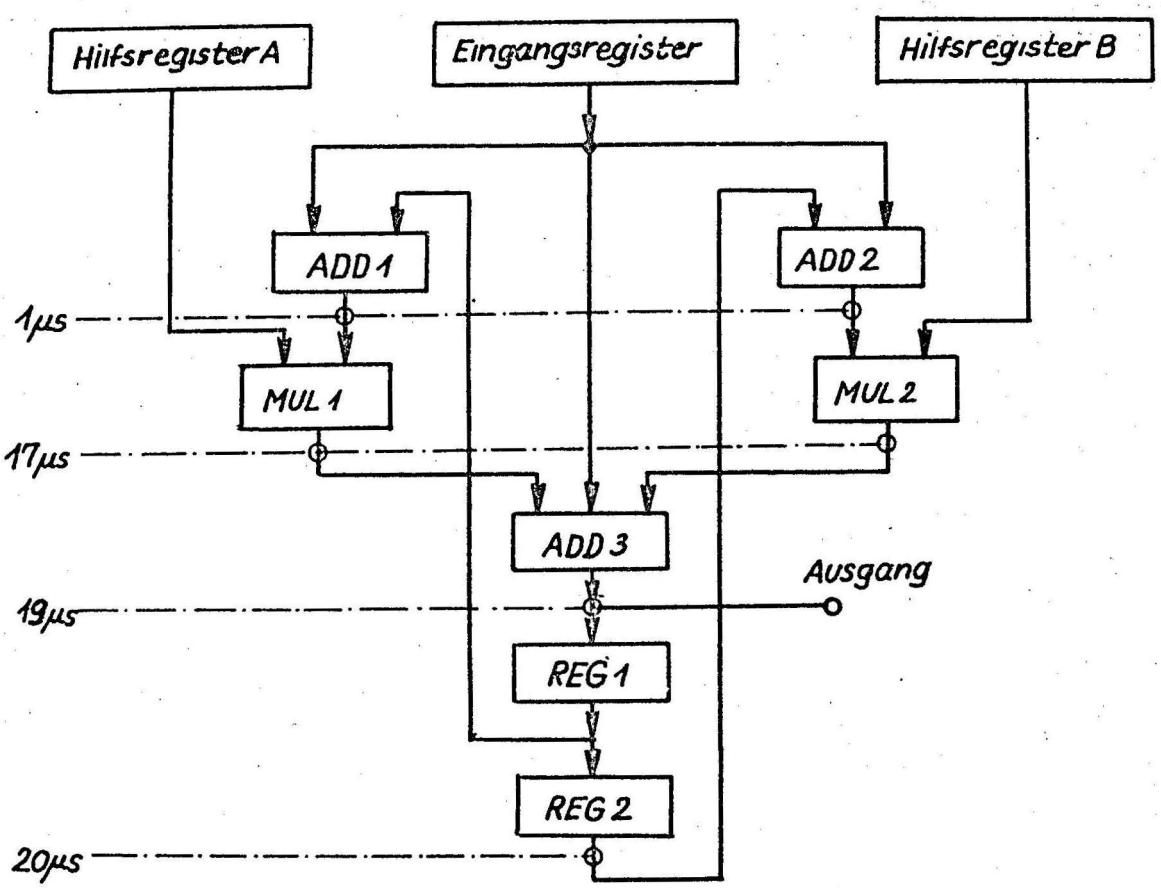


Abb.90, Schematischer Ablauf der Rechenoperationen

beiden parallelen Multiplizierer MUL1 und MUL2 die Summen ADD1 und ADD2 uebernehmen und weiterverarbeiten. Die 16 bit-Multiplizierer benoetigen ca 16 µs Rechenzeit, so dass nach 17 µs der Addierer ADD3 das erste Produkt von MUL1 zum Inhalt des Eingangsregisters hinzuaddieren kann. Nach 18 µs

wird das Produkt von MUL2 noch hinzuaddiert. Nach insgesamt 20 µs steht das Ergebnis fest und die Inhalte der Register REG1 und REG2 sind weitergeschiftet worden.

In Abb.91 ist ein vollstaendiges Blockschaltbild des Formantfilters dargestellt. Der zentrale Takt wird von einem quarzgesteuerten Taktgenerator mit der Frequenz 10.0 MHz erzeugt. Der Takt wird durch einen 200-Zaehler auf 50 kHz heruntergeteilt. Aus dem 200-Zaehler werden durch eine Decodierschaltung alle Zeitpunkte herausgegriffen, die zur Steuerung des einzelnen Formantnetzwerkes benoetigt werden.

Der 50 kHz-Takt am Ausgang des 200-Zaehlers wird durch einen 5-Zaehler auf 10 kHz herabgesetzt. Der 5-Zaehler steuert den Multiplexer fuer das Formantfilter.

Zur Darstellung von fuenf verschiedenen Formanten muessen 3 × 2 variable Koeffizienten fuer die ersten drei Formanten und 2 × 2 feste Koeffizienten fuer den vierten und fuenften Formanten den Multiplizierern des Formantnetzwerkes zur Verfuegung gestellt werden. Ausserdem muessen 5 × 2 Inhalte der Register REG1 und REG2 ausgetauscht werden. Die Zwischenspeicherung der 16 variablen Werte (3 × 2 Koeffizienten und 5 × 2 Zeitfunktionswerte) wird in einem 16 × 16 bit Flipflopspeicher durchgefuehrt. Die Adressierung dieses Speichers wird ueber eine Decodierung ebenfalls vom 5-Zaehler bewerkstelligt.

Die 3 × 2 variablen Koeffizienten fuer die ersten drei Formanten werden nach der Erzeugung von 100 (oder 200) Zeitfunktionswerten entsprechend 10 ms (oder 20 ms) Sprache vom Digitalrechner durch neue ersetzt. Das richtige Einschreiben in den Flipflopspeicher wird durch eine Steuerschaltung bewerkstelligt, die jeweils nach 100 (oder 200) Durchlaeufen des 5-Zaehlers gestartet wird.

Eine Schwierigkeit der Reihenschaltung von Formantgliedern bestand darin, die einzelnen Formanten richtig auszusteuern. Aus diesem Grunde werden die Eingangswerte der Formantschaltung so geschiftet, dass die signifikanten Stellen unmittelbar hinter dem Vorzeichenbit erscheinen. Die Anzahl der Schiftschritte wird in einem gesonderten Zaehler abgespeichert und dem Signal erst unmittelbar vor der Digital-Analog-Umsetzung am Ausgang wieder zugesetzt.

Die Amplitudenkontrolle MA stellt eine Multiplikation mit Potenzen von 2 dar. Ihr Wert wird dem Ausgangsregister des Formantsynthetisators ebenfalls durch eine entsprechende Anzahl von Schiftschritten uebertragen.

Das Ausgangssignal des Digital-Analog-Umsetzers wird durch einen Tiefpass geglaettet und ueber einen Lautsprecher abgestrahlt.

Der beschriebene digitale Formantsynthetisator wird z.Z. auf dem Rechner IBM 360/67 simuliert und getestet.

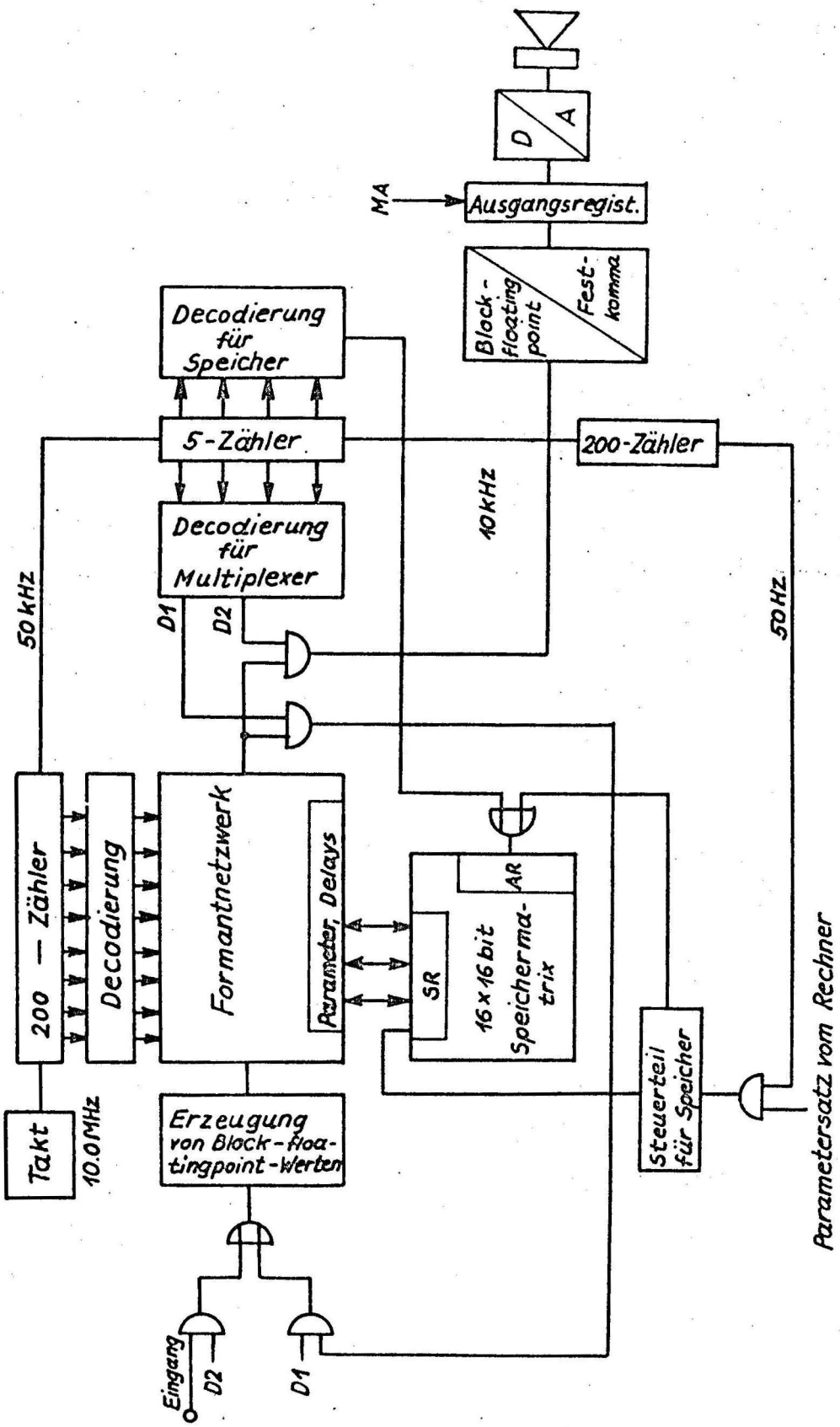


Abb.91, Blockschaltbild des Formantfilters

Verzeichnis der verwendeten Literatur

/1/ J.L.Flanagan, C.H.Coker, L.R.Rabiner, R.W.Schafer, N.Umeda
Synthetic Voices for Computers
IEEE Spectrum, Okt.1970

/2/ J.L.Flanagan
Speech Analysis Synthesis and Perception
Springer Verlag Berlin, Heidelberg, New York, 1965

/3/ T.Takasugi and J.Suzuki
Speculation of Glottal Waveform from Speech Wave
Journal of the Radio Research Laboratories
Japan, Nov.1968

/4/ B.Gold and C.M.Rader
Systems for Compressing the Bandwidth of Speech
IEEE Trans. on Audio and Electroacoustics, Vol.AU-15,3
Sept. 1967

/5/ R.M.Gold
Digital Computer Simulation of a Sampled- Data Voice-Excited Vocoder
Jour. Acoust. Soc. Amer. Vol.35, No.9
Sept.1963

/6/ J.L.Flanagan and R.M.Golden
Phase Vocoder
The Bell System Technical Journal
Nov. 1966

/7/ M.R.Schroeder
Vocoders: Analysis and Synthesis of Speech
Proceedings of the IEEE
May 1966

/8/ F.Winckel
Der Vocoder - codierte Uebertragung und Verarbeitung von Sprache
ETZ Bd.19, H.22, 1967

/9/ C.-E.Liedtke
Technischer Bericht Nr.114 des Heinrich Hertz Instituts fuer Schwingungsforschung
Berlin, 1970

/10/ W.Giloi, M.Krause, C.-E.Liedtke
Computergesteuerte Spracherzeugung
4. Kybernetik-Kongress, Berlin, 1970

/11/ J.A.Howard, R.C.Wood
Hybrid Simulation of Speech Waveforms Utilizing a Gaussian Wave Function Representation
Simulation, Sept.1968

/12/ D.R.Reddy
Pitch Period Determination of Speech Sounds
Communication of the ACM, Vol.10, No.6
June 1967

/13/ B.Hafemeister
Sprachanalyse nach Gaussschen Funktionen
Studienarbeit am Institut fuer Informantionsverarbeitung I der Technischen Universitaet Berlin, 1969
Betreuer: Liedtke

/14/ B.Hafemeister
Sprachsynthese aus Gaussschen Funktionen
Diplomarbeit am Institut fuer Informationsverarbeitung I der Technischen Universitaet Berlin, 1970
Betreuer: Liedtke

/15/ C.-E.Liedtke
Anwendung von Hybridrechnern fuer Sprachverarbeitung
AICA-IFIP Conference on Hybrid Computation
Muenchen, 1970

/16/ A.M.Noll
Cepstrum Pitch Determination
Jour. Acoust. Soc. Amer. 41, 1967

/17/ Cooley - Tukey
An Algorithm for the Machine Calculation of Complex Fourier Series
Mathematics of Computation, Vol.19,
April 1965

/18/ W.Kuntz, W.Schuessler, W.Winkelnkemper
Untersuchungen ueber Eigenschaften, Entwurf und Realisierung digitaler Filter
Erlangen, 1969

/19/ H.M.Christiansen, L.Schweizer
New Correlation Vocoder
Jour. Acoust. Soc. Amer., 1966

/20/ E.Knapp
Kanalvocoder
Diplomarbeit am Institut fuer Informationsverarbeitung I der Technischen Universitaet Berlin, 1969
Betreuer: Liedtke

/21/ B.Gold and C.M.Rader
Digital Processing of Signals
McGraw- Hill Book Company, 1969

/22/ L.K.Schweizer
Problems in Realising a Digital Vocoder, and Novel Solutions
International Seminar on Digital Processing of Analog Signals, Zurich, 1970

/23/ L.R.Rabiner
Digital Formant Synthesizer for Speech Synthesis Studies
Jour. Acoust. Soc. Amer. Vol.43, No.4, 1968

/24/ G.Fant et al.
OVE II Synthesis Strategy
Speech Communication Seminar, Stockholm, 1962

/25/ B.Gold and C.M.Rabiner
The Channel Vocoder
IEEE Trans. on Audio and Electroacoustics, Vol.AU-15,4 Dec. 1967

/26/ B.Betzenhammer
Ein Formantvocoder nach dem Frequenzzaehlverfahren
Telefunken- Zeitung, Jg.40, Heft 1/2, 1967

/27/ DeClerk, Phyfe, Fisch
Formant Synthesizer Electronically Controlled
Speech Communication Seminar, Stockholm, 1962

/28/ J.Liljencrants
The OVE III Speech Synthesizer
IEEE Trans. on Audio and Electroacoustics, Vol.AU-16, 1 March 1968

/29/ J.Anthony
A Resonance Analogue Speech Synthesizer
Fourth International Congress of Acoustics
Kopenhagen, 1962

/30/ T.Hirasaki, H.Date, H.Nakajima
A Computer Controlled Terminal Analog Speech Synthesizer
The 6-th International Congress on Acoustics
Tokyo, 1968

/31/ R.Meisenburg
Magnetbandmarkierung
Studienarbeit am Institut fuer Informationsverarbeitung I der Technischen Universitaet Berlin, 1969
Betreuer: Liedtke

/32/ G.Gerull
Visible Speech
Studienarbeit am Institut fuer Informationsverarbeitung I der Technischen Universitaet Berlin, 1969
Betreuer: Liedtke

/33/ W.Noffz
Visible Speech II
Studienarbeit am Institut fuer Informationsverarbeitung I der Technischen Universitaet Berlin, 1970
Betreuer: Liedtke

/34/ A.G.Deczky
Recursive Digital Filter Synthesis Using Fourier Series
3. Kolloquium Digitale Systeme fuer Filterung und Simulation, Universitaet Erlangen, 1969

/35/ A.V.Oppenheim, R.W.Schafer, T.G.Stockham Jr.
Nonlinear Filtering of Multiplied and Convolved Signals
Proceedings of the IEEE, Vol.56, No.8, August, 1968

/36/ R.Kossak
Schnelle Pitchbestimmung
Diplomarbeit am Institut fuer Informationsverarbeitung I der Technischen Universitaet Berlin, 1970
Betreuer: Liedtke

/37/ J.Saniter
Schnelle Cepstral-Analyse
Diplomarbeit am Institut fuer Informationsverarbeitung I der Technischen Universitaet Berlin, 1970
Betreuer: Liedtke

/38/ O.Tulgan
Kurzzeit-Cepstrum
Studienarbeit am Institut fuer Informationsverarbeitung I der Technischen Universitaet Berlin, 1969
Betreuer: Liedtke

/39/ Nakatsin und Suzuki
Formant Frequency Extraction Using Inverse Filtering and Moment Calculation and Its Evaluation by Synthesized Speech
Journal of the Radio Research Laboratories, Vol.16, 83 Japan, 1969

/40/ R.Schafer, R.Rabiner
System for Automatic Formant Analysis of Voiced Speech
Jour. Acoust. Soc. Amer., Vol.47, No.2, 1970

/41/ C.-E.Liedtke
Pole- Zero- Determination of the Vocal-Tract-Transfer Function
International Seminar on Digital Processing of Analog Signals, Zurich, 1970

/42/ Kraft
Formant Analyse
Studienarbeit am Institut fuer Informationsverarbeitung I der Technischen Universitaet Berlin, =969
Betreuer: Liedtke

/43/ Vormelcher
Automatische Formantbestimmung
Diplomarbeit am Institut fuer Informationsverarbeitung I der Technischen Universitaet Berlin, 1970
Betreuer: Liedtke

/44/ C.-E.Liedtke
Entwicklung eines Rauschgenerators
Studienarbeit am Institut fuer Informationsverarbeitung I der Technischen Universitaet Berlin, 1966

/45/ Bronzite
Audio- Spectrum- Analyzer
Electronic Engineering, Jan 1968

/46/ Gold and Rabiner
Analysis of Digital and Analog Formant Synthesizers
IEEE Trans. on Audio and Electroacoustics, Vol.AU-16, 1 March 1968

/47/ Kraft
Analoger Formantvocoder
Diplomarbeit am Institut fuer Informationsverarbeitung I der Technischen Universitaet Berlin, 1970
Betreuer: Liedtke