ARBEITEN ZUR SPRACHSYNTHESE AN DER TU DRESDEN

Ulrich Kordon, Heiko Petzold, Antje Wirth

Technische Universität Dresden

Die Arbeiten zur Sprachsynthese an der TU Dresden wurden 1968 aufgenommen. Ziel der Aktivitäten war die Realisierung eines phonemgesteuerten Formantsynthetisators als Labormuster.1977 konnte mit dem Sprachsynthetisator ROSY ein 4-Formant-Parallelfiltersynthetisator fertiggestellt werden, der von einem Kleinrechnersystem angesteuert und in einem Fahrplanauskunftssystem eingesetzt wurde. Die Ansteuerung erfolgte auf Phonembasis bei entsprechender Kontextanalyse durch das Steuersystem des Synthetisators.

Nach einer Unterbrechung von ca. 8 Jahren fanden diese Arbeiten ihre Fortsetzung. Mit der zur Verfügung stehenden Kapazität war eine Einengung der Themenbreite unerläßlich. Deshalb rückte der Anwendungsaspekt in den Vordergrund. Entsprechend gesellschaftlicher Erfordernisse richtete sich das Hauptinteresse auf Behindertenhilfsmittel mit Sprachausgabe, insbesondere für Blinde bzw. hochgradig Sehbehinderte.

Die Orientierung auf anwendungsorientierte Forschung bedeutete jedoch nicht den Verzicht auf eigene Arbeiten zu den technischen Komponenten der Sprachausgabesysteme bzw. dem zugehörigen phonetischen Umfeld. Ziel dieses Beitrages ist es, die diesbezüglichen Vorstellungen und Ergebnisse zusammenzufassen, als Basis für tiefergreifende Kontakte zu Einrichtungen in Industrie und Forschung, die sich mit ähnlichen Problemen befassen und an einer nutzbringenden Kooperation interessiert sind.

1. Hardware-Basis

Der gewählte Einsatzbereich erfordert Synthesesysteme, die sowohl freie Texte beliebiger Länge als auch Äußerungen von wenigen, festen Worten generieren können. Um dabei eine möglichst gute Sprachqualität zu erreichen, wurde auf zwei unterschiedliche Sprachsyntheseverfahren orientiert: die klassische Formantsynthese für zeichengesteuerte Synthese bzw. ein Sprachwiedergabeverfahren, ähnlich CVSD, für begrenzte Wortschätze mit festem Diskursbereich.

Für Anwendungsfälle mit geringeren Anforderungen an die Sprachqualität wurde als low-cost-Version der Sprachsynthetisator TUSY2 vorgesehen [1] . Das Anregungssystem besteht aus einem Pseudo-Zufallsgenerator als Rauschquelle und einem Rechteckgenerator als Impulsquelle. Über eine Schaltlogik wird das jeweils erforderliche Anregungssignal an 4 in Reihe geschaltete Formantfilter (Integratorschleifen, [2]) angelegt, die in Mittenfrequenz und Bandbreite variierbar sind. Als Amplitudensteuerung ist ein schaltbarer Spannungsteiler vorgesehen. Die Parametervariation der Filter erfolgt über umschaltbare Widerstandskombinationen. Für die Ansteuerung sind 4 Steuerbytes für ein Zeitfenster erforderlich. Die einstellbaren Parameterbereiche sind in Tab. 1 angegeben. Der Synthetisator ist mit dem integrierten Synthetisator MEA 8000 vergleichbar. Die diskrete Realisierung wurde gewählt um Untersuchungen zu optimalen Parameterbereichen vornehmen zu können.

Für leistungsfähigere Synthesesysteme wurde der Synthetisator TUSY3 [1] entwickelt. Als Anregungssystem kommt hier ein digitaler Funktionsgenerator zum Einsatz, der verschiedene stimmhaft- und stimmlos-Anregungsfunktionen liefern kann. Das Formantfiltersystem besteht ebenfalls aus vier Integratorschleifen, deren Bandbreiten- und Mittenfrequenzsteuerung allerdings über multiplizierende Digital-Analog-Wandler vorgenommen wird. Parallel zu diesen Reihenfilterzweig wurde ein nur bei stimmlosen Signalen angeregter Filterkanal angeordnet, der drei weitere in Reihe geschaltete und entsprechend in Mittenfrequenz und Bandbreite steuerbare Filter umfaßt.

Damit können zwei Frikativ (FF)- und ein Antiformant(AF) realisiert werden. Beide Filterzweige sind getrennt über geschaltete Spannungsteiler in der Amplitude steuerbar. Zur Ansteuerung sind 16 Byte je Zeitfenster erforderlich. Tab. 1 gibt die realisierbaren Parameterbereiche an.

	Anregung	Filtertrakt	Paralleltrakt	Amplitude
TUSY2	stimmhaft/ stimmlos Grundfrequenz fG: 85-150 Hz	f1: 150-997 Hz B1: 40-700 Hz f2: 490-2962 Hz B2: 120-700 Hz f3:1059-2919 Hz B3: 140-700 Hz f4: 3500 Hz B4: 140-700 Hz		A: 0..-27dB
TUSY3	stimmhaft/ stimmlos Grundfrequenz fG: 87-153 H Amplitude stl- Anregung AFa: max/min versch. Stimmhaft-Funktion.	f1: 16-3953 Hz f2: 16-3953 Hz f3: 32-8160 Hz f4: 32-8160 Hz B1,B2,B3,B4: 0,005-0,39 relativ	fAF:35-8925 Hz BAF:0,03-0,55 relativ fFF1:35- 8925 Hz BFF1:0,005- 0,39 relativ fFF2:35- 8925 HZ BFF2:0,005- 0,39 relativ	Vokal- trakt AV: 0..-36dB Parallel- trakt AF: 0..-29dB

Tab. 1 Einstellbare Parameterbereiche der Formantsynthetisatoren TUSY2/3

Die Realisierung des Anregungssystems in Form eines digitalten Funktionsgenerators [3] erlaubt die Erzeugung verschiedener Anregungssignalformen, die bei der Aufbereitung der Ansteuerdaten entsprechend maximaler Sprachqualität ausgewählt, im Synthesefall zeitfensterabhängig variiert werden. Momentan sind vier unterschiedliche stimmhaft- und stimmlos-Signale abrufbar. (Bild 1)

Ziel gegenwärtiger Arbeiten ist die Realisierung eines CAFS- (Codebuch-angeregtes Formant-Synthese)-Systems. Ähnlich den CELP-Systemen erfolgt im Synthesedatenaufbereitungsprozeß die Bestimmung eines Codebuches für die auftretenden, im Sinne einer hohen Sprachqualität, erforderlichen Anregungssignaltypen. Die Ansteuerdaten beinhalten dann zusätzlich den Code

der jeweils zu aktivierenden Anregung. Ein Blockschaltbild der Formantsynthetisatoren TUSY2 und 3 zeigt Bild 2.

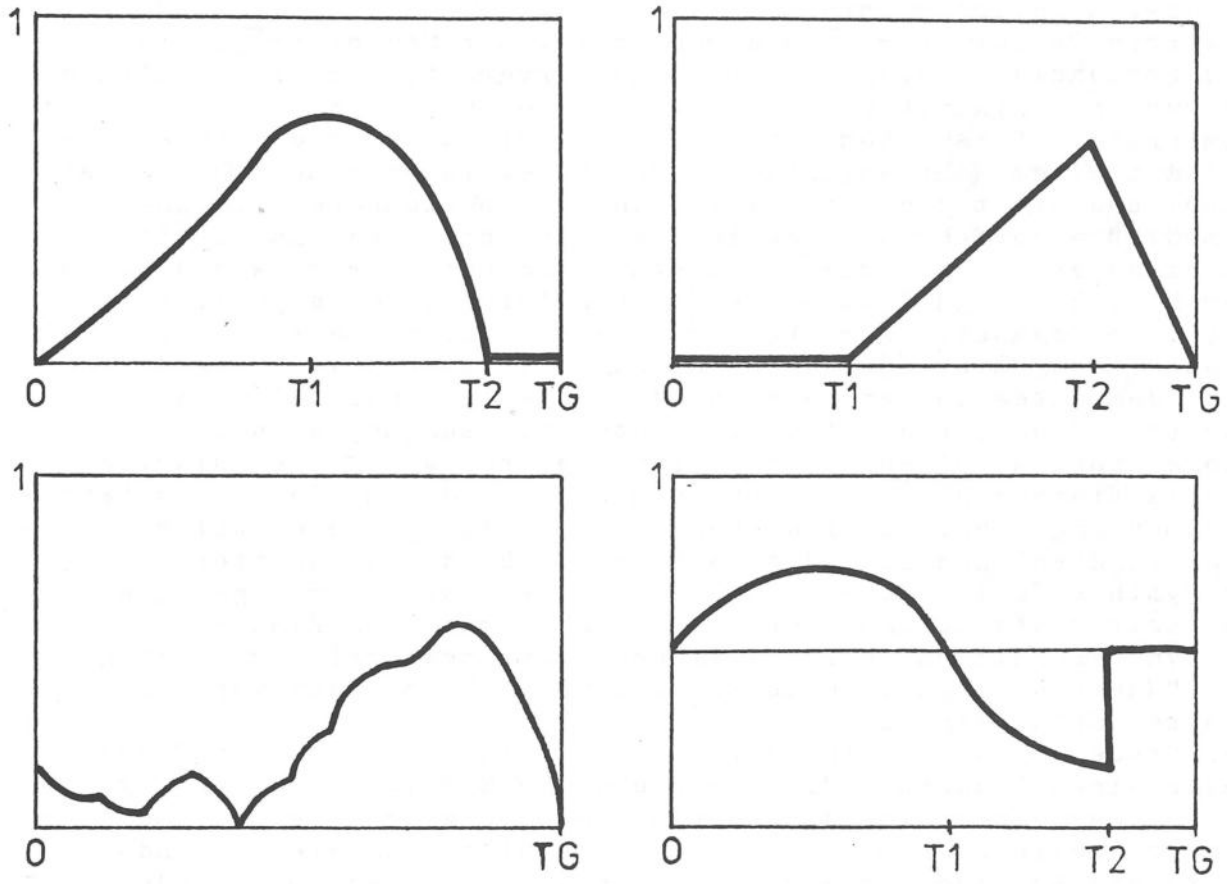

Bild 1 Stimmhaft-Anregungssignaltypen (nach [4]), TG = 1/fg

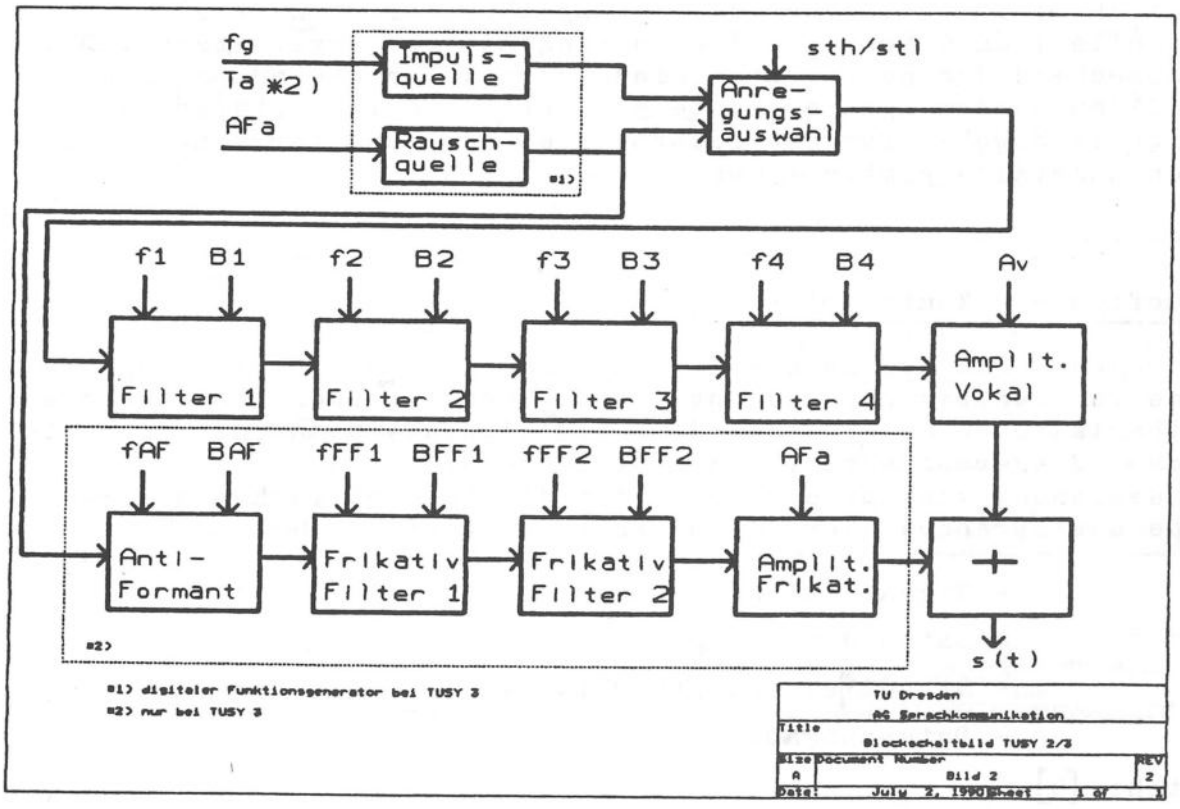

Bild 2 Blockschaltbild der Formantsynthetisatoren TUSY2/3

Für den Einsatz in Sprachwiedergabesystem ist das "Codec-G"-Verfahren entwickelt worden [5]. Prinzipiell ist es mit dem CVSD - Verfahren vergleichbar, erfordert jedoch keinen Multiplizierer und ist weitestgehend resistent gegen Bitfehler im digitalisierten Sprachsignal.
Das Prinzip ist in Bild 3 dargestellt. Zur Codierung gelangt das Sprachsignal über die üblichen Vorverarbeitungsstufen (Bandbegrenzung, Dynamikregelung, evtl. Preemphasis) an einen Komparator. Dieser Komparator vergleicht das zu codierende Signal mit einem Näherungssignal. Das Ergebnis wird zwischengespeichert und steht damit zur weiteren Verarbeitung bereit. Analog zum CVSD - Verfahren detektiert eineentsprechende Logik vier aufeinanderfolgende gleiche Zustände im Datenstrom, was jeweils durch High - Signal am Eingang eines Integrators signalisiert wird. Im Gegensatz zum CVSD - Verfahren dient das Ausgangssignal dieses Integrators jedoch nicht zur multiplikativen Verknüpfung mit dem integrierten Datenstrom zum Näherungssignal am Komparator. Vielmehr dient dieses Integratorausgangssignal als Eingangssignal eines weiteren Integrators, wobei die Polarität dieses Eingangssignals in Abhängigkeit von der Bitfolge des Datenstromes umgeschaltet wird. Das Ausgangssignal dieseszweiten Integrators dient nun als Näherungssignal für den Komparator.
Im Synthesefall wird der in einem Festwertspeicher abgelegte Datenstrom des zu generierenden Sprachsignals am Eingang der im Synthesefall nur erforderlichen Näherungsschaltung angelegt. Das "Näherungssignal" entspricht dann praktisch dem reproduzierten Sprachsignal.
Zur Steuerung der Synthesesysteme ist ein Steuermodul auf der Basis eines Einchip - Mikrorechners (UB 8830, ähnl. Z8) vorgesehen, der einerseits die Kopplung an übergeordnete Systeme über eine serielle oder parallele Schnittstelle realisiert , andererseits die Ansteuerung der Synthetisatoren mit entsprechenden Synthesedaten übernimmt (Bild 4).
Bei den Formantsynthetisatoren ist dabei sowohl reproduktive als auch minimalzeichengesteuerte Synthese (Diphonbasis) einschließlich Graphem - Ansteuerung möglich (vgl. Abschn. 3). Entsprechend dem geforderten Einsatzzweck können aus diesen Grundkomponenten Synthesesysteme unterschiedlicher Leistungsfähigkeit sowohl für den Laborbetrieb als auch für praktische Anwendungsfälle kombiniertwerden.

2. Software - Komponenten

Die zugehörigen Software - Komponenten umfassen Entwicklungssysteme zur Generierung und Optimierung von Ansteuerdaten für die Synthetisatoren aus realen Sprachsignalen sowie anwendungsspezifische Steuersoftware für den Steuermodul.
Zur Gewinnung von Ansteuerdaten für die Formantsynthetisatoren wurde das Spracheditiersystem SPREDI entwickelt, das aus

- Signalanalyse
- Dateneditierung
- Grundfrequenzaufbereitung
- Datenausgabe

besteht [6].

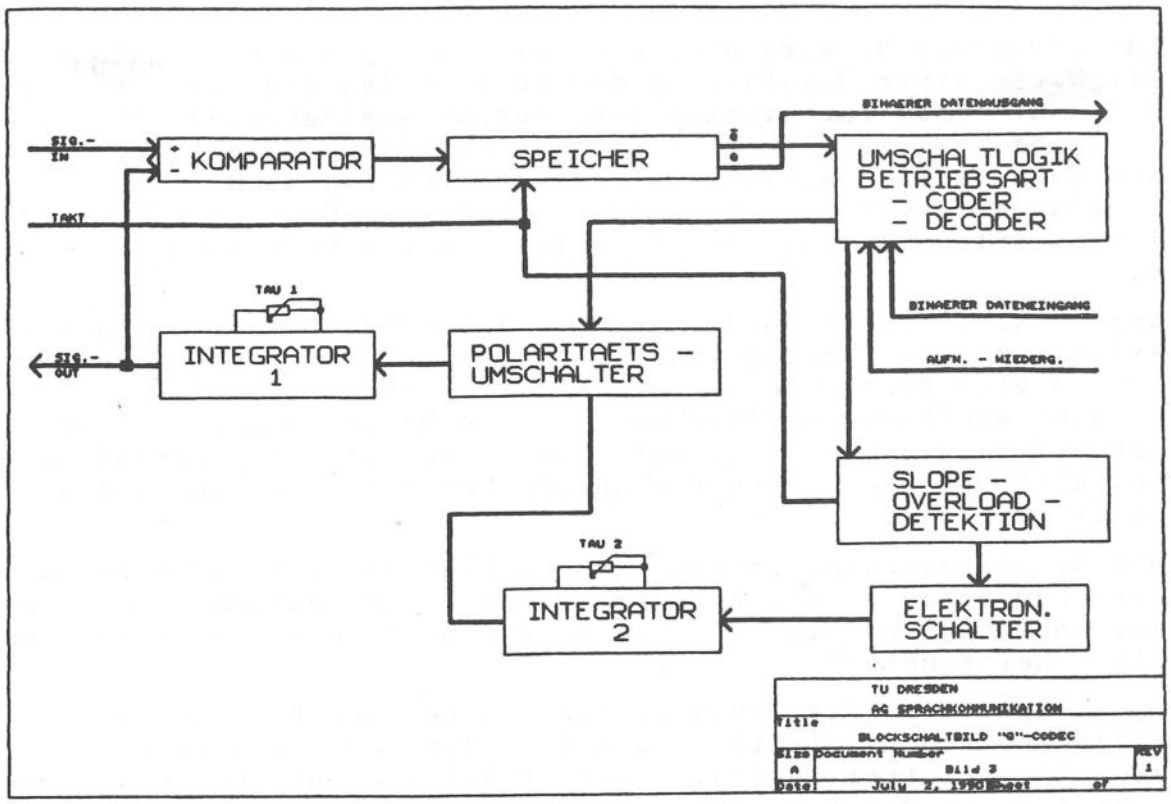

Bild 3 Blockschaltbild des "Codec-G"-Systems

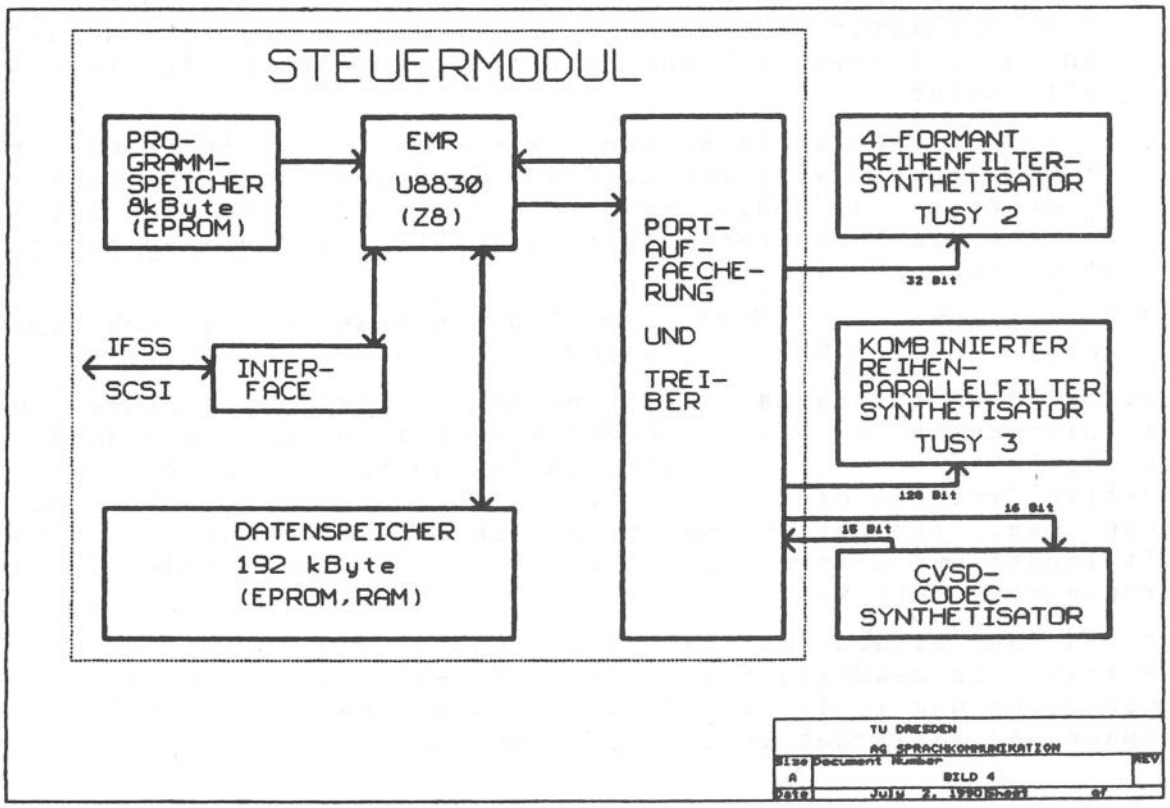

Bild 4 Blockschaltbild des Steuermoduls

Das Sprachsignal wird über ein12bit-ADC-System digitalisiert. Mit Hilfe einer Darstellung der Energie des eingelesenen Signals auf einem Farbdisplay kann der zu analysierende Abschnitt ausgewählt werden. Außerdem ist es möglich die Zeitfunktionen des selektierten Abschnitts darzustellen und nach DA-Rückwandlung akustisch zu kontrollieren. Länge und Überlappung bei Analyse aufeinanderfolgender Signalabschnitte können variiert werden.

Nach HAMMING-Fensterbewertung des aktuellen Signalabschnitts erfolgt die Bestimmung von Mittenfrequenz und Bandbreite der ersten vier Formanten aus einem LPC-Spektrum (LPC-Koeffizientenzahl wählbar) des Abschnitts durch Maximasuche. Die berechneten Parameter-Zeit-Verläufe können auf dem Farbdisplay dargestellt und über einen Formantsynthetisator zur Rücksynthese genutzt werden.

Die Grundfrequenzberechnung kann wahlweise mittels AMDF- oder Cepstrumverfahren durchgeführt werden. Die Bestimmung der Gesamtamplitude erfolgt aus dem Betragsmittelwert der Abtastwerte eines Zeitfensters.

Durch die Dateneditierung sollen die bei der Analyse erzeugten Rohdatensätze hinsichtlich Qualität des damit erzeugten Synthesesignals optimiert werden. Gestützt auf die optische und akustische Ausgabe der Datensätze können interaktiv folgende Manipulationen ausgeführt werden:

a) Variation von Parameterwerten wie Änderung der Werte für Mittenfrequenzen, Bandbreiten, Gesamtsignalamplitude, Grundfrequenz in einem Zeitfenster oder in einem Bereich, der sich über mehrere Zeitfenster erstreckt; Anhebung oder Absenkung aller innerhalb eines Bereiches liegender Werte für einen Parameter um einen beliebig wählbaren Wert; Vertauschen von Formantverläufen innerhalb eines Bereiches oder in einem Zeitfenster.

b) Manipulationen am Parameter-Zeit-Verlauf ohne Änderung der Parameterwerte wie: Vertauschen, Einfügen, Herausnehmen, Verdoppeln und Kombinieren einzelner Abschnitte oder Zeitfenster des Parameterverlaufes, Verkürzen, Verlängern der Steuerdatensätze.

c) Generierung von "künstlichen" Datensätzen (ohne vorherige Analyse realer Sprachsignale).

Bei der Grundfrequenzaufbereitung ist es möglich, Konturen der Mikrointonation von Konsonant-Vokal-Verbindungen (nach MEHNERT) in die Parameter-Zeitverläufe von Datenmaterial für die reproduktive Synthese einzuarbeiten. Für Verbindungen von Nasalen, Liquiden, stimmhaften und stimmlosen Explosiv- und Engelauten mit langen und kurzen Vokalen liegen die entsprechenden Grundfrequenzverläufe vor.

Es kann aus mindestens zwei Verläufen für einen Lautübergang gewählt, die gewählte Kontur durch Verschiebung, Streckung oder Stauchung in den vorliegenden Parameter-Zeitverlauf eingepaßt und das Ergebnis synthetisiert werden.

Die Dateiausgabe gestattet das

a) Erstellen von Dateien, in denen alle von den genutzten Synthetisatoren realisierbaren Parameterwerte enthalten sind

b) Erstellen von EPROM-Dateien, in denen die Parameter-Zeit-Verläufe in auf die zu nutzende Synthesehardware angepaßter Form vorliegen

c) Erstellen von Dateien, die Konturen der Mikrointonation von Lautübergängen enthalten.

Das System SPREDI ist in der Programmiersprache FORTRAN 77 realisiert und ist auf einem Rechner vom Typ K1630 (DEC-kompatibel) lauffähig. Neben den Synthetisatoren TUSY2/3 können damit Daten für den Schaltkreis MEA 8000 generiert werden.

Die Aufbereitung von Daten für das "Codec-G"-System ist mit dem Programm CGEDI möglich [7].

Folgende Grundfunktionen können damit ausgeführt werden:

a) Codierung von Sprachsignalen über "Codec-G"-Coder

b) Datenaustausch und -speicherung mit dem Rechner des Aufbereitungssystems

c) Editieren: bitweises Modifizieren der Datensätze
Anfangs-/Endpunktbestimmung
Löschen von Datensätzen (bzw. -teilen)
Kombination von Datensätzen
Generierung einer EPROM-Datei

d) Reproduktion von Sprachsignalen aus Datensätzen über "Codec-G"-Decoder. Das System ist in Turbo-Pascal realisiert und auf einem AC A7150 (PC-kompatibel) lauffähig.

Die Software des Steuermoduls für die Synthetisatoren ist abhängig vom Anwendungsfall. Für die im Abschnitt 4 angegebenen konkreten Objekte wurden Versionen erarbeitet.

3. Phonetisches Umfeld

Im Vordergrund stehen hier Untersuchungen zu einem graphemzeichengesteuerten Synthesesystem. Schwerpunkte bilden dabei ein Graphem-Phonem-Umsetzalgorithmus sowie die Entwicklung eines Diphon-Systems für die Formantsynthetisatoren. Die Graphem-Phonem-Umsetzung [8] arbeitet regelgestützt und arbeitet dreistufig. In einer ersten Stufe werden Ganzworte wie Eigennamen, Sonderfälle und häufig wiederkehrende Worte vollständig oder in Teilen umgesetzt. Die zweite Stufe nimmt eine Vorphonemisierung vor. Im Ergebnis dieses Verarbeitungsschrittes liegt bereits eine vollständige Phonemfolge als Ergebnis vor. Lediglich die relativ komplizierten Vokale erhalten einen Zwischencode. Mit der dritten Verarbeitungsstufe werden auch diese Zwischencodes in den endgültigen Phonemcode umgesetzt. Neben der Umsetzung von Alphazeichen erlaubt das Graphem-Phonem-Umsetzsystem GRAPHO die Phonemisierung von Sonderzeichen und Zahlen einschließlich Datumsangaben, Uhrzeiten und Ordnungszahlen. Das Programm GRAPHO ist in Assemblersprache für den Prozessor UB8830 des Steuermoduls realisiert. Durch eine besondere Speicherorganisation des Regelwerkspeichers ergeben sich relativ kurze Umsetzzeiten von wenigen Millisekunden/Wort.

Dem Diphon-System DIPHOS liegen 41 Einzellaute zugrunde, aus denen ca. 1600 mögliche Lautverbindungen erzeugt werden können. Aufgrund von Erkenntnissen, die im Laufe der Arbeiten an der Erstellung solcher Diphonsteuerdatenfolgen gewonnen wurden, konnte diese Zahl auf ca. 600 Diphone verringert werden. Einige der dadurch nicht explizit abgespeicherten Lautverbindungen werden mit Hilfe von Regeln erzeugt. Dazu gehören alle Verbindungen, bei denen auf einen beliebigen Laut ein Explosivlaut folgt. Andere Lautverbindungen können durch einfaches Aneinanderreihen der Diphone erzeugt werden. Hierzu gehören alle Verbindungen von Lauten mit folgenden stimmhaften Engelauten, die wenig vom vorhergehenden Kontext abhängen.

Das Diphon-System DIPHOS befindet sich gegenwärtig noch in der Entwicklung. Die Implementierung erfolgt ebenfalls auf dem im Abschnitt 1 erwähnten Steuermodul.

4. Anwendungen

Die in Abschnitt 1 bis 3 vorgestellten Ergebnisse grundlegender Untersuchungen wurden in verschiedenen Blindenhilfsmitteln genutzt. So entstanden ein wissenschaftlicher Taschenrechner mit Sprachausgabe, ein Sprachausgabemodul als Korrekturhilfe für Schreibsysteme sowie verschiedene Applikationen des "Codec-G"-Verfahrens (z.B. elektronisches "Notizbuch").

Im Beitrag "Sprachsynthese-Anwendungen" in diesem Band werden diese Geräte näher vorgestellt.

5. Literatur

/1/ Petzold,H.: Technische Prinzipien zur Sprachsynthese. Dissertation A, TU Dresden, in Vorbereitung

/2/ DD; WP 2380872; Int. Cl.: WP H 03K. Kaskadierbares aktives Filter zur Erzeugung synthetischer Sprachlaute. Naumburger,V.; Jäger, G.; Widemuth, E., 30.11.1983

/3/ Kordon, U.: Anregungssystem für Formantsynthetisatoren auf der Basis eines digitalen Funktionsgenerators. In: Studientexte zur Sprachkommunikation, Heft 7, TU Dresden 1990, S. 56-63

/4/ Quach-Tuan, N.; Guerin, B.: Voice Excitation Sources for Digital Formantsynthesizer. Bulletin du Laboratoire de la Communication Parlee No 1 A 1987, S. 67-89

/5/ Grunitz, O.: Codierverfahren für Sprachspeicher. Diplomarbeit, TU Dresden, Sektion Informationstechnik 1989

/6/ Kordon, U.; Stein, A.; Petzold, H.: Ein System für die Gewinnung von Daten für die Sprachsynthese. In: Studientexte zur Sprachkommunikation, Heft 4, TU Dresden 1987, S. 47-52

/7/ Römke, H.: Sprachsynthesemodule für PC. Großer Beleg TU Dresden, Sektion Informationstechnik 1990

/8/ Heymann, J.: Graphem-Phonem-Umsetzung mit einem Ein-Chip-Mikrorechner. Diplomarbeit, TU Dresden, Sektion Informationstechnik 1989