Sprachsignalerkennung und Sprachsynthese

Dr.-Ing. ULRICH KORDON

Mitteilung aus der Sektion Informationstechnik der TU Dresden

Die Erkennung und die Synthese von Sprachsignalen haben durch die Fortschritte auf dem Gebiet der Mikroelektronik und der damit möglichen Interpretation komplizierter Systeme zunehmende praktische Bedeutung erlangt. Dieser Beitrag beschäftigt sich mit den heute üblichen Prinzipien derartiger Systeme und stellt anhand ausgewählter Beispiele den erreichten Stand dar.

In den letzten Jahren sind auf dem Gebiet der Verarbeitung von Sprachsignalen durch technische Systeme wesentliche Fortschritte erreicht worden. So konnten Einrichtungen zur Erzeugung von Sprache im größeren Rahmen marktwirksam werden. Auch Geräte zur Sprachsignalerkennung fanden erste kommerzielle Anwendung. Charakteristische Eigenschaften dieser zu einer 1. Generation von Sprachverarbeitungssystemen zu rechnenden Lösungen sind

- Einzelwortbasis (z. B. Erkennung bzw. Erzeugung von Befehlsworten, max. kurzen Wortfolgen)
- bei Erkennungssystemen Sprecherabhängigkeit; d. h. die für die Erkennung notwendigen Bezugssprachmuster müssen in einem Lernprozeß für jeden Sprecher individuell generiert werden
- begrenztes Wortinventar (technisch und konzeptionell bedingt).

Trotz dieser einschränkenden Randbedingungen stießen derartige Systeme durch ihre Anpassungsfähigkeit an verschiedene Einsatzfälle bei einem vertretbaren Preis-Leistungsverhältnis auf entsprechendes Interesse. So wurden für 1983 und 1984 Zuwachsraten auf diesem Gebiet um 400 % angegeben [1].
Motiviert durch die Überlegungen, nach denen für Computer der 5. Rechnergeneration Interfaces für den Informationsaustausch mit Hilfe der menschlichen Sprache ein wichtiges Merkmal darstellen werden, wurden die Forschungen auf den Gebieten der Spracherkennung und Sprachsynthese weiter forciert. Auch der prinzipielle Trend zu technischen Systemen allgemein mit exponierten „intelligenten" Eigenschaften trug zu dieser Entwicklung bei.
Die Ergebnisse zur Reduzierung der Sprecherabhängigkeit und der Verarbeitungsmöglichkeit fließender Sprache bei großen Wortschätzen stellen eine wichtige Grundlage zur Realisierung entsprechend leistungsfähigerer Spracherkennungssysteme dar. Bei der Sprachsynthese konzentrieren sich die Arbeiten auf eine Steigerung der Natürlichkeit der synthetischen Sprache und die Vervollkommnung der Ansteuerbedingungen (z. B. Schriftzeichensteuerung, automatische Textgenerierung).
Als wesentlichste Voraussetzung für das Erreichen dieses Niveaus, sowohl bei den praktisch bereits nutzbaren Systemen als auch im Forschungsbereich, sind jedoch die mit der Entwicklung der Mikroelektronik verbundenen technologischen Fortschritte anzusehen. Ökonomisch günstige Schaltkreislösungen bildeten einerseits die Grundlage für die breitere Einführung von Sprachtechnik in die Praxis und sind andererseits Voraussetzung für die technische Absicherung anspruchsvoller Grundlagenuntersuchungen. Der Anwendungsbereich erstreckt sich dabei vom Konsumgütersektor (z. B. sprechende Uhren, Waagen, Tanksäulen, sprachgesteuertes Spielzeug bzw. Geräte der Heimelektronik u. a.) über Sprachein- und Sprachausgabeports für Personal-, Büro- und Heimcomputer bis hin zu Experimentiersystemen in der Forschung. Unter Umständen können mit dem Einsatz von Spracherkennungs- bzw. Sprachsynthesesystemen Probleme bewältigt werden, für deren Lösung die konventionelle Technik keine oder wesentlich ungünstigere Alternativen bietet, z. B. bei Extrembedingungen wie Dunkelheit, Feuchte, Schmutz oder in der Rehabilitationstechnik (Blindenhilfsmittel u. a.).
Ziel dieses Beitrages ist es, anhand einiger typischer Lösungen den momentan erreichten technologischen Stand auf den Gebieten der Spracherkennung und Sprachsynthese aufzuzeigen. Im Vordergrund sollen dabei vor allem Konzepte unter Nutzung moderner Spezialschaltkreise für die Sprachsignalverarbeitung stehen.

Sprachsynthese

Die zur Sprachsignalerzeugung verfügbaren Systeme unterscheiden sich zunächst im verwendeten Steuerungsprinzip, d. h. im Niveau der Information, die vom steuernden System zur Auslösung eines bestimmten Synthesevorgangs bereitgestellt werden muß. Für einfache Lösungen geringer Universalität mit exakt abgestecktem Einsatzbereich kommt fast ausnahmslos das Verfahren der reproduktiven Sprachsynthese zur Anwendung (Bild 1).
Der Synthetisator besteht aus einem Digitalspeicher (ROM, EPROM), einem Dekoder sowie einem Digital-Analog-Umsetzer (DAU) mit nachfolgendem Tiefpaß und NF-Verstärker. Zur Steuerung ist eine entsprechende Logik vorhanden. Beim Hersteller werden alle später im Einsatz zu synthetisierenden Worte bzw. Wortfolgen von einem Sprecher gesprochen, die entsprechenden analogen Sprachsignale in digitale gewandelt und nach einer redundanzmindernden Kodierung im Digitalspeicher des Synthetisators abgelegt.
In jüngster Zeit ist dieser Vorgang mit Hilfe sogenannter Spracheditiersysteme auch vom Anwender derartiger Synthetisatoren durchzuführen, womit eine höhere Universalität erreicht wurde. Bei der Synthese ist nun lediglich eine Speicheradresse als Steuerinformation an den Synthetisator anzulegen. Das ab dieser Adresse abgespeicherte „Wort" bzw. die abgespeicherte „Wortfolge" wird über Dekoder und DAU wieder in ein analoges Sprachsignal zurückgewandelt. Dabei muß mit einem Datenfluß von etwa 1 000 bit/s gerechnet werden. Verständlichkeit und Natürlichkeit der synthetischen Sprache sind ausgezeichnet, jedoch ist nur ein kleines fest vorgegebenes Vokabular bei trotzdem erheblichem Speicherbedarf realisierbar. Bei merklich eingeschränkter Natürlichkeit des Synthesesprachsignals kann die Universalität der Synthetisatoren durch Ausnutzung der minimalzeichengesteuerten Sprachsynthese wesentlich erhöht werden (Bild 2).
Hier ist eine Programmierung des Synthesesystems für alle später zu synthetisierenden Worte oder Wortfolgen nicht notwendig. Der Hersteller speichert lediglich laut- bzw. lautteiltypische Signalkonfigurationen als Parametersätze kodiert in einem Digitalspeicher ab. Außerdem sind Parameterübergangsverläufe zwischen Einzellauten bzw. Lautteilen

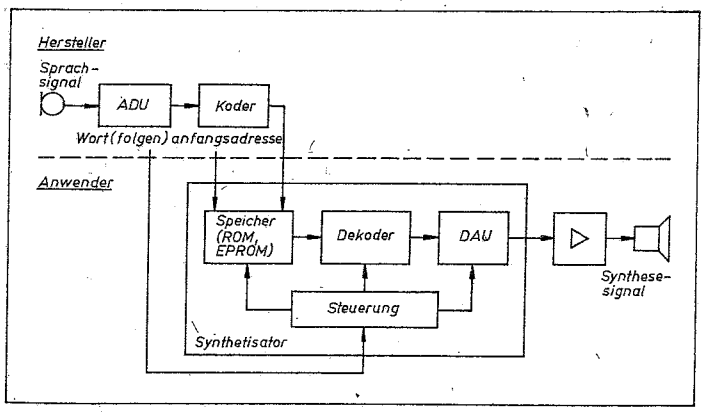

Bild 1: Reproduktive Sprachsynthese

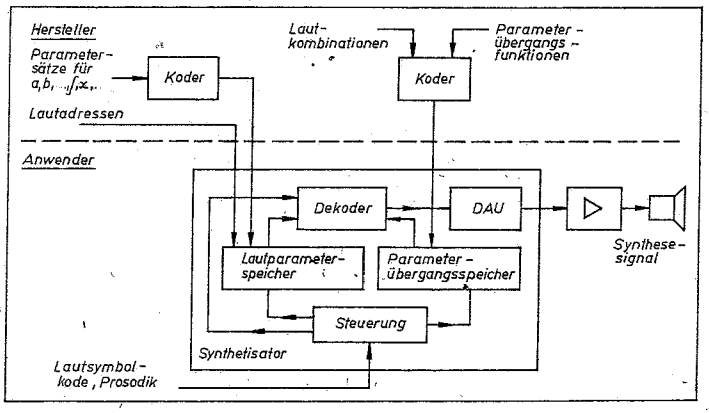

Bild 2: Minimalzeichengesteuerte Sprachsynthese

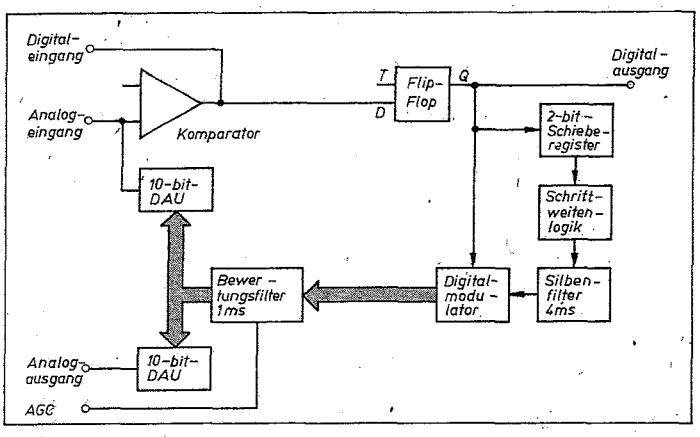

Bild 3: Prinzipschaltung des Sprachsynthetisators HC-55564 (Harris) [11]

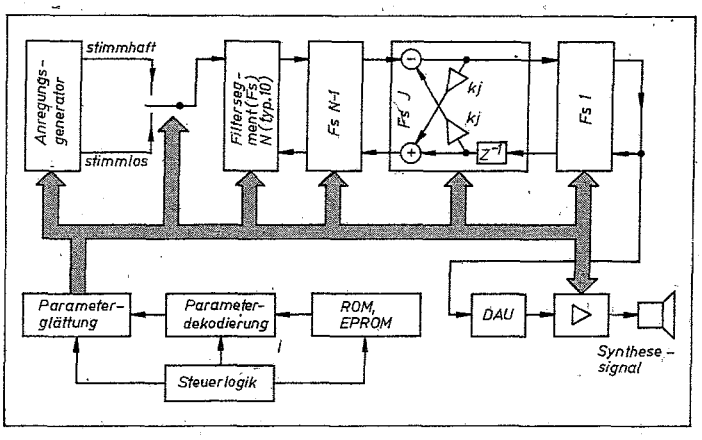

Bild 4: Sprachsynthetisator auf LPC-Basis

sowie lautkontextabhängige Modifikationen der Parameter gespeichert. Im Syntheseeinsatz braucht der Nutzer nur die Folge der zu synthetisierenden Laute in Form einer Lautkodefolge sowie evtl. zusätzlich entsprechende prosodische Lautinformationen (wie Lautdauer, Lautgrundfrequenz und Lautpegel) an den Synthetisator zu übergeben. Eine umfangreiche Steuerung reiht die zu Signalabschnitten dekodierten entsprechenden Lautparameter gemäß der gewünschten Lautfolge aneinander und gibt das Signal aus. Es ist ersichtlich, daß hiermit auch nahezu alle beliebigen Fremdsprachen erzeugt werden können. Aufgrund der hohen Abstraktion ist nur ein Datenfluß von etwa 70 bit/s zu bewältigen. Synthetisatoren mit einer solchen Steuerung finden zum Beispiel Anwendung in Vorleseautomaten für Sehbehinderte o. ä., da hier reproduktive Verfahren nur sehr bedingt einsetzbar sind.

Für die Kodierung der Sprachsignale bzw. Dekodierung im Synthetisator haben sich im wesentlichen drei Grundverfahren durchgesetzt [2]:

- Adaptive Deltamodulationsverfahren, einschließlich modifizierter Techniken wie die CVSD-Kodierung (Continuously Variable Slope-Delta) und das speziell für Sprachsignale entwickelte Mozerverfahren
- Linear-Prediction-Kodierung (LPC)
- Formantverfahren.

Prinzipiell muß festgestellt werden, daß Steuerprinzip und Kodierung nicht unabhängig voneinander eingesetzt werden können. So erfordert ein universelles Steuerungskonzept wie die Minimalzeichensteuerung eine entsprechend abstrahierende Kodierung. Für die reproduktive Synthese wäre dagegen bei genügend großem Speicher die direkte Nutzung der digitalisierten Abtastwerte des zu synthetisierenden Sprachsignals, d. h. praktisch unkodiert, als Syntheseparameter möglich.

Beim CVSD-Verfahren wird der Kodierer durch einen Quantisierer gebildet, der die Differenzen zwischen dem zu kodierenden Sprachsignal und einem Näherungssignal getaktet in Bitfolgen umsetzt. Diese Bitfolgen werden unmittelbar als Sprachdaten gespeichert. Ein H-Pegel entspricht dabei einer positiven Differenz, L-Pegel einer negativen. Typisch für das CVSD-Verfahren ist die Bildung des Näherungssignals. Prinzipiell wird dazu die Ausgangsbitfolge integriert. Damit sich auch kurzzeitige Änderungen des zu kodierenden Sprachsignals in der Bitfolge niederschlagen, darf das Näherungssignal keine allzu großen Differenzen zu diesem aufweisen. Deshalb wird das Ausgangssignal des Koders multiplikativ mit einem Steuersignal verknüpft. Die Größe des Steuersignals hängt dabei von der Anzahl aufeinanderfolgender gleicher Zustände der Bitfolge am Koderausgang ab. Bei größeren Differenzen wird das Näherungssignal somit entsprechend verstärkt und damit schneller nachgeführt. Der mit Speicher und Steuerung den Synthetisator bildende Dekoder realisiert die umgekehrte Funktion. Die Systemstruktur eines typischen Syntheseschaltkreises für reproduktive Synthese auf CVSD-Basis zeigt Bild 3. Bemerkenswert ist die Möglichkeit für den Nutzer, durch Umschaltung auf Kodieren selbst Sprachdaten generieren zu können. Für 4 s Sprache werden etwa 8 Kbyte Speicher benötigt.

Bei LPC-Synthetisatoren (Bild 4) bilden in Analogie zum humanen Spracherzeugungssystem ein zwischen Rausch- und Pulsfolgesignal umschaltbarer Anregungsgenerator und ein mit zeitlich veränderlichen Koeffizienten gesteuertes Digitalfilter als Artikulationstrakt-Modell den Dekoder gemäß Bild 1 bzw. Bild 2. Das Anregungssignal soll durch das steuerbare Digitalfilter so verformt werden, daß es einem Sprachsignal entspricht. Bei zu erzeugenden dynamischen Signalen (z. B. Worte, Wortfolgen) wird das Signal als Folge kurzzeitig stationärer Abschnitte aufgefaßt, wobei im Kodierungsprozeß die für die Synthese des jeweiligen Abschnittes erforderlichen Filterkoeffizienten aus natürlicher Sprache ermittelt werden. Diese Folge von Koeffizientensätzen wird mit der außerdem für jeden Abschnitt bestimmten Anregungssignalinformation (stimmlos, Pulsabstand), Signalamplitude sowie evtl. weiteren Kodierungsdaten (z. B. Anzahl der Wiederholungen dieses Parametersatzes) im Speicher abgelegt. Werden diese Informationen lautorientiert übermittelt, ist bei entsprechender Steuerung diese Kodierung neben reproduktiven auch in lautgesteuerten Synthetisatoren einsetzbar. Das Synthesefilter ist in der Regel eine Gitterstruktur mit zwei Multiplizierern, zwei Summierern und einem Speicher. Der Analysealgorithmus des Koders zur Ermittlung der Filterkoeffizienten bei vorgegebenem Sprachsignal ist ausführlich in [3] dargestellt. Die Koeffizienten besitzen 3 bis 5 bit Auflösung, wobei für die Synthese stimmhafter Abschnitte etwa zehn, bei stimmlosen etwa vier notwendig sind (Abtastfrequenz etwa 8 kHz, obere Signalfrequenz 4 kHz). Für Amplitude und Pulsfolgedauer stehen je 4 bis 5 bit sowie ein Bit für die Wiederholungsinformation zur Verfügung. Der DAU besitzt eine Auflösung von 8 bit. Moderne Strukturen schließen einen Speicher von 20 bis 30 Kbit (extern erweiterbar) sowie eine entsprechende Steuerlogik mit ein. Einfachere Schaltkreise besitzen Schnittstellen zum Anschluß externer Speicher (z. B. über 128 bit FIFO-Speicher, mehrere 12 Kbit ROM) und Steuerprozessoren (z. B. 8 bit Daten, READY, INT). Bei Verwendung von pMOS- bzw. nMOS-Technologien und Betriebsspannungen zwischen 5 V und 9 V beträgt die Leistungsaufnahme 70...200 mW.

Im Unterschied zu LPC-Synthetisatoren, bei denen die spektrale Formung des neutralen Anregungssignals durch ein komplexes steuerbares Filter höherer Ordnung realisiert wird, besteht bei Formantsynthetisatoren dieses Bewertungssystem aus etwa drei bis fünf steuerbaren Einzelfiltern. Diese sind entsprechend den als Formanten bezeichneten informationstragenden spektralen Maxima des Sprachsignals in Mittenfrequenz, Bandbreite und evtl. in der Amplitude steuerbar. Wegen der Analogien zum menschlichen Artikulationstrakt hat sich die Reihenschaltung dieser Filter durchgesetzt. Diese Reihenfilteranordnung bildet mit dem wie bei LPC-Synthetisatoren verwendeten Anregungssystem den Dekoder. Die Formantkodierung ist sowohl in reproduktiven als auch minimalzeichengesteuerten Systemen anwendbar. Beim Kodierungsprozeß werden die Parameter der spektralen Maxima, wie Gesamtamplitude und Stimmlos- bzw. Grundfrequenzinformation, aus dem natürlichen Sprachsignal bestimmt und als Syntheseparameter abgespeichert. Eine vor allem in Personalcomputern verwendete Einchip-Lösung zur Sprachsynthese auf Formantbasis ist der Schaltkreis SSI 263 von Silicon Systems Inc. [4]. Es handelt sich dabei um einen lautzeichengesteuerten Synthetisator, d. h., der integrierte Sprachparameterspeicher enthält bereits alle erforderlichen Formantdaten und entsprechende Lautübergangsinformationen. Zur Feineinstellung können diese Parameter über fünf 8-bit-Register extern variiert werden. Bild 5 zeigt die Struktur des SSI 263. Der jeweils zu synthetisierende Laut wird über einen 6-bit-Kode aktiviert, zwei Bits dienen zur Lautdauereinstellung. Der Grundfrequenzverlauf ist über weitere 8 bit in zwei Betriebsarten (4096 Stufen ≙ sieben Oktaven bzw. 32 Stufen mit acht verschiedenen Übergangsverläufen zwischen diesen Werten) steuerbar. Außerdem sind im Bedarfsfall die Ausgabegeschwindigkeit, Gesamtamplituden sowie Filter- und Übergangscharakteristika beeinflußbar. Der Filtertrakt besteht aus fünf kaskadierten Tiefpaßfiltern in SC-Technik. Bei geringeren Ansprüchen an die Sprachqualität kann auf die Möglichkeiten der Parametervariation über die fünf Register verzichtet werden. Der SSI 263 ist in 5-V-CMOS-Technologie gefertigt. Seine Inter-

facebedingungen erlauben eine problemlose Einbindung in entsprechende Mikroprozessorsysteme. Als externe Beschaltung ist lediglich ein NF-Verstärker erforderlich.

In Abhängigkeit von eventuell benutzten Feineinstellungen sind Speicher zwischen 70 bit und 500 bit zur Synthese von 1 s Sprache erforderlich. Eine Auswahl von Sprachsynthese-Schaltkreisen mit einigen wesentlichen Parametern ist in der Tafel zusammengestellt. Es kann festgestellt werden, daß für alle gegenwärtig interessanten Anwendungsfälle geeignete Schaltkreislösungen verfügbar sind. Obwohl derartige Angaben in der Literatur gewöhnlich mit Vorsicht registriert werden müssen, weisen Preise von einigen 10 US-Dollar für derartige Schaltkreise auf auch ökonomisch günstige Einsatzmöglichkeiten hin.

Der Entwicklungstrend dürfte mit dem bereits erwähnten Schaltkreis SSI 263 deutlich werden: Verbesserung der Sprachqualität durch mögliche Variation der Syntheseparameter und zunehmende Einbeziehung prosodischer Informationen (also über Metrik und Rhythmik) wie Grundfrequenzverlauf und Akzentuierung bei universeller und einfacher Nutzbarkeit durch Minimalzeichensteuerung. Die Realisierung als Peripherieschaltkreis von Mikrorechnersystemen ist als tragfähige technologische Variante anzusehen.

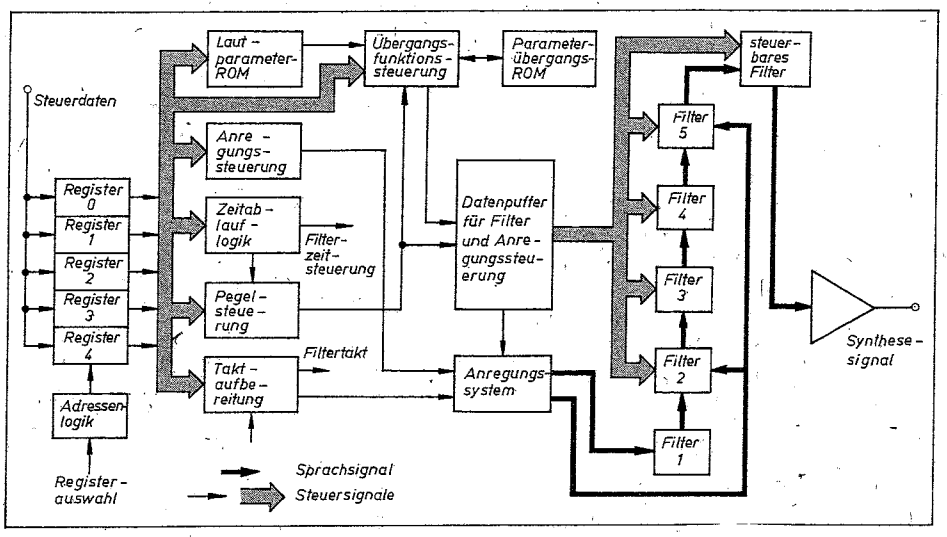

Bild 5: Prinzipschaltung des lautzeichengesteuerten Sprachsynthetisators SSI 263 (Silicon Systems Inc.) [4]

Spracherkennung

Die gegenwärtig angebotenen Spracherkennungssysteme auf der Basis integrierter Schaltungen arbeiten fast ausnahmslos sprecherabhängig auf Einzelwortniveau. Zwischen den zu erkennenden Worten sind deutliche Sprechpausen erforderlich. Der gesamte Wortschatz ist für jeden Sprecher einzeln anzulernen.

Das verwendete Funktionsprinzip entspricht weitestgehend dem Analysator-Klassifikator-Modell der allgemeinen Objekterkennung.

Für den Ausgleich unterschiedlicher zeitlicher Relationen zwischen zu erkennendem Sprachsignal und den in der Lernphase gebildeten Bezugsmustern im Referenzwissen des Erkenners sowie die dadurch bestimmte Strategie zur Ermittlung der Ähnlichkeiten zwischen ihnen kommen vor allem zwei Konzepte zur Anwendung.

Bei kleineren Spracherkennungssystemen (etwa 50 Worte) ist folgende Vorgehensweise typisch (Bild 6). Im Analysator wird das zu erkennende Sprachsignal in eine Folge gleich langer äquidistanter Abschnitte zerlegt. Aus jedem Abschnitt wird eine Anzahl relevanter Merkmale extrahiert, die als Komponenten eines diesen Abschnitt beschreibenden Vektors aufgefaßt werden können. Das Signal wird praktisch in eine Vektorfolge transformiert.

Jedes später zu erkennende Wort muß der jeweilige Nutzer in der Lernphase nun mehrmals sprechen (drei bis sechs Wortrealisierungen), wobei die vom Analysator ermittelten Vektorfolgen gespeichert werden. Durch lineare Interpolation werden nun alle Vektorfolgen auf eine einheitliche Länge (Vektorzahl) gestaucht bzw. gestreckt sowie die zu einem Wort gehörenden längennormierten Vektorfolgen zu einer als jeweiliges Bezugsmuster dienenden Referenzvektorfolge gemittelt. Die damit für jedes zu erkennende Wort vorliegenden Bezugsmuster bilden das Referenzwissen des Erkenners. In der Erkennungsphase wird die Vektorfolge des zu erkennenden Wortes ebenfalls auf die Vektorzahl des Bezugsmusters normiert, es werden die Differenzen zu allen Bezugsvektorfolgen gebildet. Der Kode des Bezugsmusters mit der geringsten Differenz zur zu erkennenden Folge wird schließlich als Erkennungsergebnis ausgegeben.

Problematisch wird die Anwendung dieses Prinzips, wenn erhebliche Differenzen in den Längen der Laute des zu erkennenden Wortes und den für die Bildung der Bezugsmuster verwendeten Worten auftreten. Dieses Problem kann durch Nutzung der dynamischen Optimierung [5] gelöst werden. Damit ist die Erkennung wesentlich größerer Wortschätze (≦ 1 000) sowie die Verarbeitung von Wortfolgen bzw. fließender Sprache mög-

Daten einiger Sprachsynthese-IS

Bezeichnung	Hersteller; Preis	Verfahren	Wortzahl; Synthesezeit in s	Speicherkapazität in bit intern	extern	Bemerkungen
HD38880B	Hitachi	Parcor	200;100	–	128 K (16 x)	
HD 61885	Hitachi	Parcor	? ; 20	32 K	128 K (16 x)	63 Texte
HC 55564	Harris	Delta	–	–	?	Datengenerierung möglich
MN 6401	Matsushita	Parcor	63;?	32 K	möglich	
MN 1261	Matsushita	Parcor		–	?	
µPD1774	NEC	Delta		48 K	?	
µPD7751C	NEC	Delta		–	?	
µPD7752C	NEC	Formant	?;27	32 K	?	
SC–01	Votrax;12$	phonemgesteuert (Formant)	–	–	–	64 Phoneme
SC–02 (SSI 263)	Silicon System Inc.	phonemgesteuert (Formant)	–	–	–	64 Phoneme
Serie II	TSI	Delta	24 bis 64;?	–	–	
Serie III	TSI	Delta	150;100	–	16 bis 128 K	
SPO 250	GI	Parcor		–	128 K	
SPO 256	GI; 46 DM	Parcor (Formant)	?;8... 20	16 K	16;32; 128 K	< 3825 Phrasen
Speech 1000	TSI	Parcor	?;200... 300	?	485 K	
TMC 0280	TI	Parcor		–	–	
TMC 0281	TI	Parcor		–	–	für 32minütige Adressen
TMS 990/306	TI; 3500 DM	Parcor	180;?	–	–	
TMS 5100	TI	Parcor	100;120	–	128 K	
TMS 5200	TI	Parcor		–	?	
MM 54104	NS	Delta	128;?	–	16 bis 128 K, max. 2 M	
MSM 5205	OKI	Delta		–	–	
S 3610	AMI	Parcor	32;?	20 K	?	
MEA 8000	Valvo/Philips	Formant	25;15	–	128 K	
UAA 1003	ITT; 39 DM	Delta	20,25;?	30 K	–	
UAA 1103	ITT	Delta	32;?	30 K	–	
UAA 1104	ITT	Parcor		?	–	
UAA 1105	ITT	Parcor		?	–	
M58817AP	Sanyo	Parcor		?	–	
VSY . 100	Sansui	Parcor		?	–	
MB 8760	Fujitsu	Parcor		?	–	
T 6721	Toshiba	Parcor		?	–	
ECL 1565	Denden	Formant		?	–	
MSM 6202RS	OKI	Delta		144 K	?	
LRN 3680	Sharp	?		32 K	?	

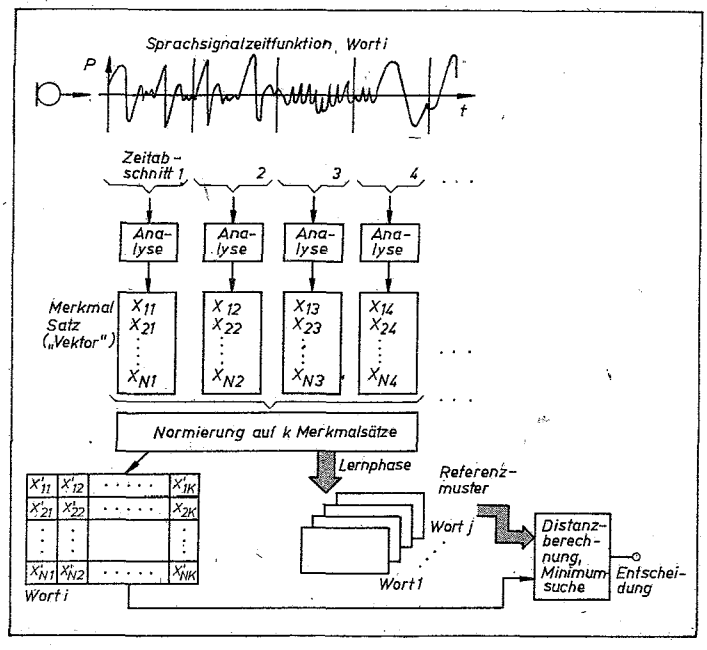

Bild 6: Prinzip einfacher Spracherkenner

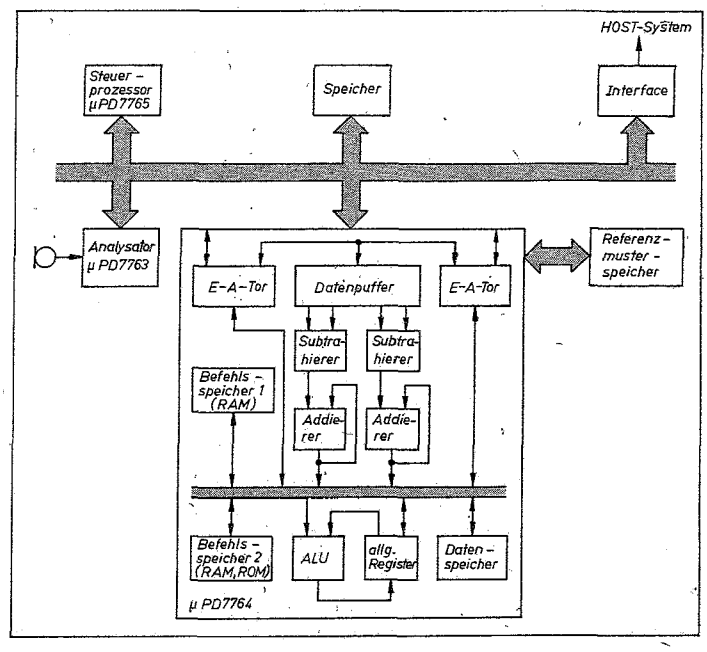

Bild 7: Prinzipschaltung der Spracherkennungs-IS µPD 7764 (Nippon Electric Corporation)

lich. Der prinzipielle Ablauf entspricht dabei dem für kleinere Erkenner. Die Interpolation zur Realisierung einer einheitlichen Länge der Vektorfolgen wird jedoch durch eine nichtlineare Abbildung der Vektorfolge des zu erkennenden Sprachsignals auf die jeweilige Bezugsmusterfolge ersetzt. Kriterium ist dabei eine minimale Distanz zwischen diesen beiden Folgen. Der Kode des Bezugsmusters, für den diese Distanz insgesamt minimal ist, kann als Erkennungsergebnis weiterverwendet werden.

Die Struktur des Erkennerschaltkreises µPD 7764 der Fa. NEC auf der Basis der dynamischen Optimierung zeigt Bild 7.

Dabei ist festzustellen, daß es sich hier um einen Bestandteil einer Bauelementefamilie handelt, d. h., zu dieser Erkenner-IS gehören noch externe Speicher, die Analysator-IS µPD 7763 sowie der Steuerprozessor µPD 7765. Im Erkennerschaltkreis wird ausschließlich die Distanzberechnung durchgeführt. Gemäß dem Algorithmus der dynamischen Optimierung müssen zwei Distanzmaße ermittelt werden: einmal die lokalen Distanzen zwischen je zwei Einzelvektoren und auf deren Basis die globale Gesamtdistanz zwischen zu erkennender Einheit und Referenzmustern. Deshalb sind in dieser nMOS-IS zwei Distanzprozessoren, nämlich der D-Prozessor zur Berechnung der lokalen Distanzen (Tschebyscheff-Distanz) und der G-Prozessor, der zwei verschiedene Möglichkeiten zur Ermittlung der globalen Distanz bietet, realisiert. Die Befehlszykluszeit beträgt 250 ns bei einem Systemtakt von 8 MHz. Zur Ermittlung einer lokalen Distanz sind 2 µs und einer globalen 30 µs erforderlich. Ein vollständiger Erkennungsvorgang dauert etwa 300 ms. Vom System können 340 Einzelworte bzw. bis zu 40 Worte umfassende Wortfolgen mit einer vom Hersteller angegebenen Erkennungsrate von 99 % erkannt werden.

Der Analysator µPD 7763 ist eine 16-Kanal-Filterbank mit Gleichrichter-Integratorteil in SC-Technik, die einen Frequenzbereich von 250...5 400 Hz in nichtlinearer Teilung erfaßt. Die Zeitfenstergröße ist zwischen 1 ms und 32 ms in sechs Stufen wählbar. Die Energiewerte an den Filterausgängen werden mit einem 8-bit-12-kHz-ADU digitalisiert und in einem 16-byte-Block als Merkmalwerte an den Erkennerschaltkreis übertragen. Außerdem kann der Analysator durch einen zwischen 0 und 46,5 dB programmierbaren Vorverstärker und einen Equalizer an die aktuellen Signalverhältnisse angepaßt werden. Der Schaltkreis ist in 5-V-CMOS-Technik realisiert.

Zur Bedienung hochwertiger Digitalarmbanduhren mit Hilfe der Sprache ist der in [6] vorgestellte Erkenner vorgesehen. Dabei ist zu bemerken, daß er Bestandteil der Uhr sein soll, d. h. für äußerst geringe Leistungsaufnahme bei extrem niedrigen Betriebsspannungen konzipiert wurde.

Der Analysator besteht aus einer siebenkanaligen Parallelfilterbank mit nachfolgenden Gleichrichter- und Integratorstufen und einstellbaren Schwellwertschaltern. Das zu erkennende Signal gelangt über einen ab einem bestimmten Schwellwert aktiven Verstärker mit frequenzabhängiger Verstärkung auf die Filter. Diese bestehen aus SC-Filtern 4. Ordnung (2/2), die den Frequenzbereich von 190...4800 Hz überstreichen. Dabei sind insgesamt nur drei Kanäle (sechs Filter 2. Ordnung) realisiert, die nächsten Kanäle bzw. der letzte Kanal werden durch die gleichen Filter, allerdings mit dem auf jeweils ein Viertel verringerten Takt gebildet. Die Signale an den Filterausgängen werden gleichgerichtet und integriert. Nach dem Durchlaufen der Schwellwertschalter (mit pegelabhängigen Schwellen) liegt aller 10 ms ein sieben binäre Komponenten umfassender Merkmalvektor vor. Die Vektoren gelangen nun in einen auch auf dem Chip integrierten, mikroprogrammgesteuerten Spezialprozessor. Dieser normiert zunächst die Länge der die zu erkennenden bzw. anzulernenden Worte beschreibenden Vektorfolgen auf 20 Vektoren. Vorher wurden die Folgen bereits reduziert, indem maximal sechs gleiche aufeinanderfolgende Vektoren zugelassen sowie nur einzeln auftretende Vektoren gestrichen wurden.

Die Kodierung der Vektorfolgen erfolgt so, daß mit drei Bits die Vektorkomponentennummer und mit je fünf Bits die Nummer des Vektors angegeben wird, ab bzw. bis zu dem die entsprechende Komponente 1 ist. Daran schließen sich Distanzberechnung und Entscheidung an. Der als Klassifikator arbeitende 8-bit-Prozessor wird von einem Sequenzer (256 × 6-Mikroprogramm-PLA), einem (512 × 21)-ROM zur Programmspeicherung, einem CMOS-RAM von 1 Kbit als Referenzmusterspeicher sowie einer ALU mit dem Grundoperationen OR, EXOR, NAND, INVERSION und ADDITION (2 × 8) ergänzt. Es sind die direkte und die indirekte Adressierung möglich.

Weiterhin sind acht Register und ein Befehlsdekoder vorhanden. Die maximale Programmlänge beträgt 180 Befehle. In einer vereinfachten Lernphase wird durch den jeweiligen Nutzer das dann residente Vokabular der zehn Zahlworte sowie der Steuerbefehle WATCH, ALARM, TIMER, CHRONO, HOMETIME angelernt.

Schließlich sei darauf hingewiesen, daß mit dem Einchip-Spracherkenner VRC 008 (vorprogrammierter Einchip-Mikrorechner 6805) der Fa. Interstate ein System verfügbar ist, das 16 Einzelworte sprecherunabhängig erkennen kann. Dabei soll eine Erkennungsrate von 90 % erreicht werden. Als nMOS- und als CMOS-Version erhältlich, bieten sich damit hervorragende Möglichkeiten für den Einsatz im Konsumgüterbereich.

Im Gegensatz zur Synthese steht der breite Einsatz von Spracherkennungstechnik erst am Anfang. Das von allen einschlägigen Herstellern geforderte „kooperative" Nutzerverhalten sowie die Anfälligkeit gegen äußere Einflüsse (z. B. Umweltgeräusche) stellen noch erhebliche Hindernisse dar. Eine komplexe Lösung dieser Probleme erfordert weitere umfangreiche Grundlagenforschung, die sich einerseits auf die Verbesserung von Teilkomponenten (vor allem Merkmalgewinnung im Analysator und Einbeziehung linguistischer Kenntnisse im Klassifikationsprozeß) richtet. Andererseits ist bereits absehbar, daß auch bei den Überlegungen zur Grundkonzeption wahrscheinlich völlig neuartige Wege beschritten werden müssen [7], um die von der Praxis geforderten Bedingungen erfüllen zu können.

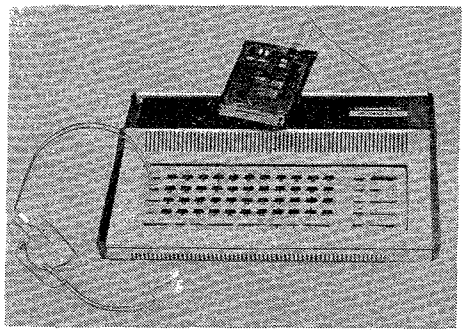

Bild 8: Entwicklungsmuster des Spracheingabemoduls für Kleincomputer KC 85/1 (TU Dresden und VEB Robotron-Elektronik)

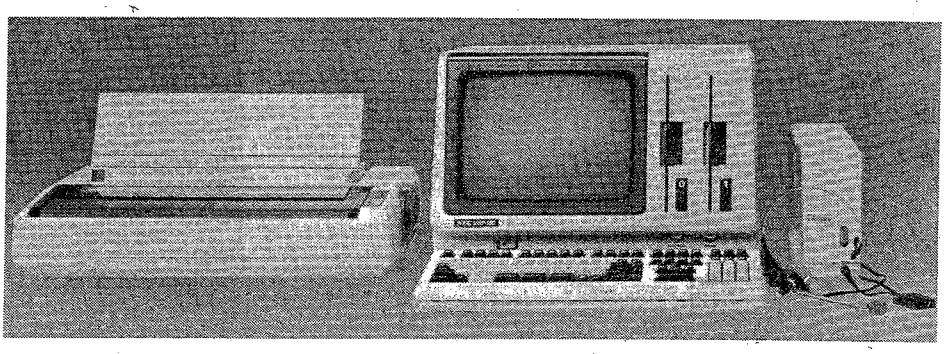

Bild 9: Spracherkenner VWP 103 N (Nippon Electric Corporation)

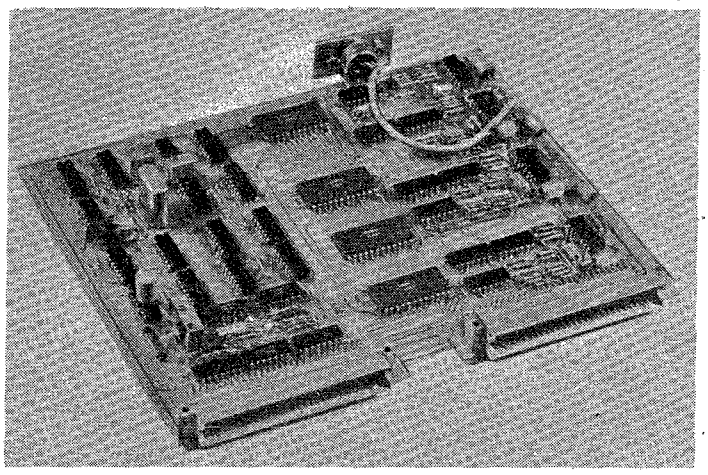

Bild 10: Sprachsynthesemodul für K 1520 auf Formantbasis (TU Dresden)

Stand in der DDR

In der DDR wird vor allem vom VEB Kombinat Robotron und der TU Dresden in Zusammenarbeit mit anderen Kooperationspartnern an der Entwicklung von Sprachein- und Sprachausgabetechnik gearbeitet. Entsprechend den geforderten Stückzahlen stehen dabei Lösungen unter Verwendung von Standardschaltkreisen im Vordergrund.

Für die sprecherabhängige Erkennung von etwa 50 Einzelworten sind der Einplatinen-Spracherkenner ESE K 7821 und Nachfolgetypen geeignet. Es handelt sich dabei um eine K-1520-OEM-Baugruppe. Sie besteht neben dem U-880-Steuerprozessor aus 16 Kbyte RAM und 4 Kbyte ROM sowie dem diskret realisierten Analysator auf Nulldurchgangsbasis [8]. Über ein Parallelinterface ist das Erkennungsergebnis abnehmbar. Für geübte Sprecher ist eine Erkennungsrate von etwa 95 % erreichbar. Als Erkennungszeit werden etwa 0,2 s angegeben. Die maximale Wortlänge darf 2 s betragen. In Verbindung mit dem Bedien-Anzeige- und Steuerteil BAS K 7822 wurde auf der Basis dieses Erkenners das Spracheingabegerät SEG K 7823 entwickelt. Es stellt ein komplettes Erkennungssystem für universellen Einsatz dar. Die prinzipielle Funktionsweise entspricht der beschriebenen mit linearer Längennormierung.

Der Spracheingabemodul für den Kleincomputer KC 85/1 bzw. KC 87 (Bild 8) stellt eine Modifikation des K 7821 dar. Er enthält den Analyseteil, während für die Klassifikation auf Ressourcen des Rechners zurückgegriffen wurde. Wesentliche Parameter und die Funktion entsprechen denen des K 7821. Als Vergleich soll hier der Spracherkenner VWP 103N der Fa. NEC angeführt werden (Bild 9). Er erkennt sprecherabhängig das japanische und lateinische Alphabet, die Ziffern 0 bis 9 sowie 56 Lautverbindungen. In Verbindung mit einem Personalcomputer und Drucker wird er vom Hersteller als „Wortprozessor, mit dem durch Sprache geschrieben werden kann" angeboten. Die technische Basis dürfte der Erkennerschaltkreis μPD 7764 sein.

Zur reproduktiven und minimalzeichengesteuerten Synthese auf Formantbasis dient der an der TU Dresden entwickelte Sprachsynthesemodul [9]. Von den vier Formantfiltern sind alle in der Bandbreite und die ersten drei zusätzlich in ihrer Mittenfrequenz steuerbar. Als Anregungssignal finden eine pseudostochastische Bitfolge bei stimmlosen Signalabschnitten sowie eine im Pulsabstand einstellbare Impulsfolge für stimmhafte Signale Verwendung. Den als K-1520-Steckeinheit realisierten Synthetisator zeigt Bild 10. Außerdem ist die Realisierung einer Variante als Modul für KC 85/1 bzw. KC 87 und einer Spezialbaugruppe für den Einsatz als Blindenhilfsmittel (z. B. sprechende Taschenrechner, Schreibmaschine o. ä.) vorgesehen. Insgesamt kann eingeschätzt werden, daß der in der DDR erreichte Stand durchaus als tragfähige technische Basis für Standardanwendungen angesehen werden kann. Auch auf dem Gebiet der Sprachsignalerkennung- und Sprachsynthese wird es zunehmend darauf ankommen, die in der Forschung erreichten guten Ergebnisse auszubauen und mit den wachsenden Möglichkeiten unserer Mikroelektronik zu verbinden.

Zusammenfassung

Der erreichte Stand bei der automatischen Sprachsignalverarbeitung ist durch einen merklichen Niveauunterschied zwischen Erkennung und Synthese gekennzeichnet.

Während das Problem der Sprachsynthese als prinzipiell gelöst betrachtet werden kann, kann bei der Erkennung nur von ersten Lösungskonzepten gesprochen werden. Das findet auch in den für praktische Anwendungen verfügbaren Bauelemente- und Systemlösungen seinen Niederschlag.

Die Fortschritte der Halbleitertechnologie werden zunehmend nutzbar gemacht. Als Beispiel sei hier der Schaltkreis SP 1000 von General Instruments erwähnt [10]. Er vereint auf einem Chip einen achtstufigen LPC-Analysator sowie ein zehnstufiges LPC-Synthesefilter mit Anregungssystem. Der mit 7,50 US-Dollar angegebene Preis erlaubt eine breite Anwendung. Eine weitere wichtige Richtung dürfte der Einsatz der in diesem Beitrag nicht betrachteten allgemeinen Signalprozessoren für Sprachzwecke sein. Durch ihre Programmiermöglichkeiten können sie problemlos an die jeweiligen Einsatzbedingungen angepaßt werden.

Die gegenwärtig verfügbaren Schaltkreislösungen für die Sprachsignalverarbeitung bieten jedoch vor allem beste Möglichkeiten, einen breiten Nutzerkreis mit dieser neuen Technik vertraut zu machen und Verständnis für die wohl auch in fernerer Zukunft noch zu fordernden einschränkenden Randbedingungen beim praktischen Einsatz solcher Systeme (z. B. erhöhte Sprechdisziplin u. a.) zu wecken. Weitere Fortschritte auf dem Gebiet der Sprachein- und Sprachausgabetechnik erfordern ein wesentlich komplexeres Herangehen aller Teildisziplinen. Nur die Einbeziehung hörpsychologischer, linguistischer, ergonomischer und anderer für die Lösung des Problems von untergeordneter Bedeutung erscheinender Erkenntnisse wird die Realisierung von Systemen fördern, die alle Vorteile, die mit dieser Technik erreichbar sind, aufweisen werden.

Literatur

[1] Tscheschner, W.: Zur Entwicklung einer neuen Generation technischer Sprachkommunikationseinrichtungen. Studientexte zur Sprachkommunikation Heft 2, S. 5–26. Dresden: Technische Universität, 1986

[2] Sickert, K.: Automatische Spracheingabe und Sprachausgabe. Haar: Markt und Technik 1983

[3] Markel, J. D.; Gray, A. H.: Linear prediction of Speech. Berlin, Heidelberg, New York: Springer Verlag 1976

[4] Soskuty, O.: Phonem-Synthesizer-IC ist μP-kompatibel. Elektronik, München 33 (1984) 17, S. 45–48

[5] Iwainsky, A.: Dynamische Optimierung. Berlin: VEB Verlag Technik 1984

[6] Bui, N. C.; Monbaron, J. J.; Michel, J. G.: An Integrated Voice Recognition System. IEEE Transactions on Acoustics, Speech and Signal Processing, New York 31 (1983) 1, S. 323–329

[7] Schroeder, M. R.: Speech an Speaker Identification. Bibliotheca Phonetica 12. Basel: Karger 1985

[8] Ito, M. R.: Zero Crossing Measurements for Analysis and Recognition of Speech Sounds. IEEE Transactions on Audio and Electroacoustics, New York AU 19 (1971) 3, S. 235–242

[9] Kordon, U.: Sprachsynthesemodul für K 1520. Nachrichtentechnik · Elektronik, Berlin 37 (1987) 1, S. 31–33

[10] Vemula, R.: Single IC Can Perform Speech Recognition and Synthesis. Electronics, New York 57 (1984) 1, S. 120–122

[11] Best, S. W.: Sprachsynthese. elektronik industrie, Heidelberg 12 (1984) 6, S. 15–27

Spracherkenner-Zusatzmodul für U-880-Mikrorechner

Dr.-Ing. LOTHAR SEVEKE und
Dr.-Ing. ULRICH KORDON

Mitteilung aus dem VEB Robotron-Elektronik Dresden und der Sektion Informationstechnik der TU Dresden

Der Einsatz von Spracherkennern als neue Komponente der Rechnerperipherie erleichtert die Mensch-Maschine-Kommunikation. Die in diesem Beitrag vorgestellte Baugruppe und das Programm für U-880-Mikrorechner realisieren mit minimalem Aufwand die Erkennung von 50 isoliert gesprochenen Wörtern in Echtzeit. Dabei ist die Erkennungssicherheit bei dem Sprecher am größten, der dem Erkenner den gewünschten Wortschatz in seiner individuellen Aussprache übermittelt hat.

Mit der Anwendung der Mikrorechentechnik in immer neuen Bereichen der Volkswirtschaft entsteht das Bedürfnis, die Kommunikation mit informationsverarbeitenden Maschinen zu verbessern, d. h., sie den Gewohnheiten der zwischenmenschlichen Kommunikation und den neuen Einsatzbedingungen (Spezifik des Arbeitsplatzes, naive Nutzer) bestmöglich anzupassen. Dies wird in Ergänzung der konventionellen Tastaturen und alphanumerischen Anzeigen bisher vor allem durch grafische Ein- bzw. Ausgabemöglichkeiten realisiert. Es gibt jedoch auch Bestrebungen, die Lautsprache, das natürliche Kommunikationsmittel des Menschen, für den Informationsaustausch zu nutzen. Dazu werden sprachliche Äußerungen des Nutzers (Wörter oder kurze Wortfolgen) in Steuerinformationen für den Rechner umgewandelt bzw. werden Informationen des Rechners an den Nutzer in Lautsprache umgesetzt.

Der Ausgabekanal für Lautsprache, der Sprachsynthetisator, ist international in breitem Maße entwickelt. Neben der Erzeugung hochqualitativer Sprache für Auskunftssysteme und nachrichtentechnische Dienste werden im internationalen Maßstab billige Sprachsynthesechips, die gut verständliche, aber noch unnatürlich klingende Sprache erzeugen, auch in Konsumgüter eingebaut.

Mit der praktischen Nutzung des Spracheingabekanals wird seit einigen Jahren ebenfalls begonnen, wobei Spracherkenner die herkömmlichen Eingabemittel nicht, etwa durchgehend ablösen sollen. Die Spracheingabe unterliegt gegenüber der Kommunikation zwischen menschlichen Partnern noch einigen Einschränkungen, die aus Grenzen der ökonomisch-technischen Machbarkeit, aber auch aus fehlendem Grundlagenwissen resultieren. Die technische Unzulänglichkeit kann hier nicht wie bei der Sprachsynthese durch die hervorragende menschliche Erkennungsfähigkeit ausgeglichen werden; das zu erkennende Signal wird durch subjektiv bedingte Sprechereinflüsse sogar noch zusätzlich gestört.

Die lautsprachliche Eingabe besitzt jedoch auch beim gegenwärtigen Entwicklungsstand in ausgewählten Einsatzfällen eine Reihe spezifischer Vorteile. So ist das Sprechen möglich, während außerdem mit Händen oder Augen andere Aufgaben gelöst werden, was beispielsweise an grafischen, Mikroskopie- oder an Sortierarbeitsplätzen von Bedeutung ist. Die Spracheingabe wird durch ungünstige Licht- und Witterungsverhältnisse kaum gestört, was ihren Einsatz bei der mobilen Datenerfassung in der Landwirtschaft, im Verkehrs- und Bauwesen begünstigt. Vor der Erkennung können Sprachsignale über Telefon oder Funksprechkanäle mit vorhandenen Geräten übertragen oder auf Magnetband gespeichert werden. Das Sprechen erfolgt außerdem mit mehr Aufmerksamkeit als die Betätigung einer Tastatur, wodurch sich Routinefehler verringern lassen.

Die hier beschriebene Spracherkenner-Baugruppe kann vor allem genutzt werden, um spezielle Hand- und Fußtastaturen zu ersetzen, um Arbeitsplätze einzusparen, an denen angesagte Informationen protokolliert werden, oder um, in Kombination mit einer Tastatur, Eingaben insgesamt zu effektivieren.

Da die breite Anwendung eines Spracherkenners wesentlich von seinem Preis abhängt, wurden Algorithmen für die Signalanalyse, das Lernen und das Erkennen ent-

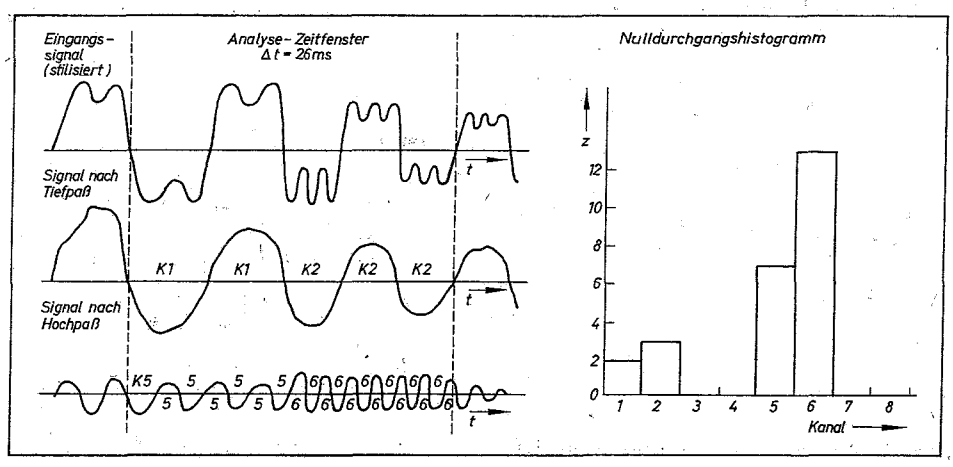

Bild 2: Bildung des Nulldurchgangshistogramms

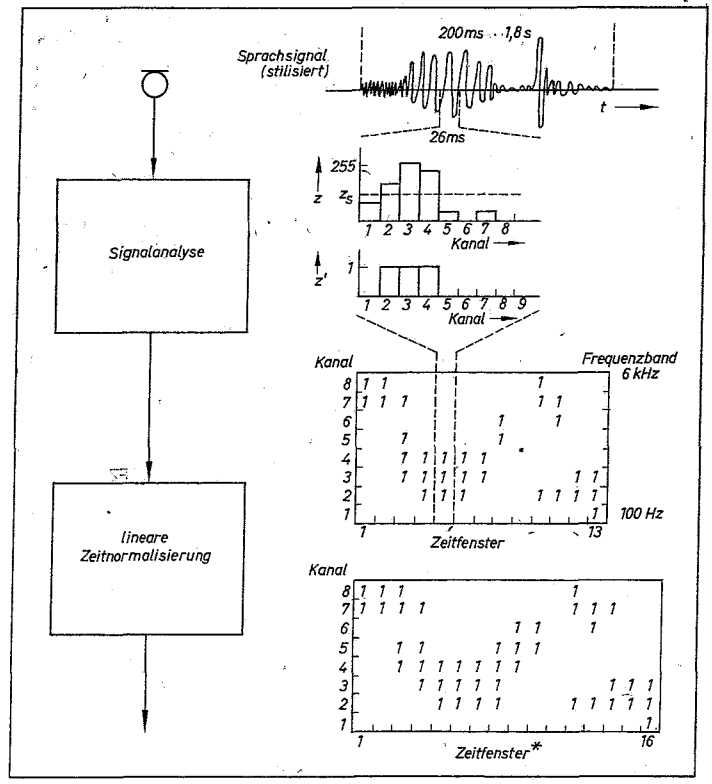

Bild 3: Bestimmung der Wortmuster

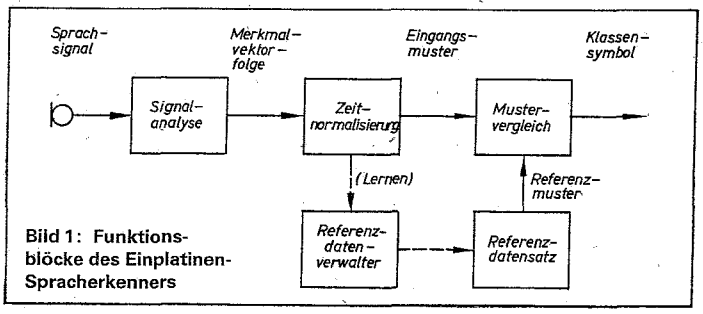

Bild 1: Funktionsblöcke des Einplatinen-Spracherkenners

wickelt, die nur eine sehr einfache Zusatzschaltung, wenig Speicherplatz im Mikrorechner und einen geringen Rechenzeitaufwand benötigen.

Funktionsweise des Worterkenners

Der im VEB Kombinat Robotron in Kooperation mit der Technischen Universität Dresden entwickelte Spracherkennerzusatz kann bis zu 50 verschiedene, isoliert gesprochene Wörter erkennen. Der Wortschatz wird durch mehrmaliges Vorsprechen (fünf- bis achtmal) im Lernprozeß vom Anwender selbst festgelegt, wobei gleichzeitig die Aussprucheeigenheiten des jeweiligen Sprechers und die Geräuschumgebung gespeichert werden. Zwischen zwei zu erkennenden Wörtern muß eine Pause von mindestens 200 ms eingehalten werden. Das Erkennungsergebnis liegt 200 ms nach dem jeweiligen Wortende in Form der im Lernvorgang vereinbarten Klassennummer vor. Die maximale Dauer einer zu erkennenden sprachlichen Äußerung beträgt 1,8 s.

Der Spracherkenner besteht aus einer Steckeinheit (Abmessungen etwa 100 mm mal 95 mm), die die Schaltung enthält und über ein CTC (U 857) die Busschnittstelle zu einem U-880-Mikrorechner realisiert, und aus einem ROM-fähigen Programm (2 Kbyte), das 4 Kbyte Arbeitsspeicher benötigt. Die wichtigsten Funktionsblöcke des Spracherkenners sind im Bild 1 dargestellt.

Signalanalyse

Im Signalanalysator wird die vom Mikrofon gelieferte Sprach-Zeit-Funktion in eine Folge von Merkmalvektoren umgewandelt, die das Signal numerisch beschreiben. Dafür wurde ein Verfahren entwickelt, das in Beachtung der Tatsache, daß absolut begrenzte Sprache noch ausreichend verständlich ist, nur die Abstände zwischen benachbarten Nulldurchgängen im Signal auswertet. Es stellt einen Kompromiß zwischen Leistungsfähigkeit und Kosten dar, da einerseits alle Informationen aus der Amplitudendynamik verlorengehen und die Informationen über Oberwellenanteile gestört sind, andererseits diese Messung mit wenig Aufwand realisierbar ist.

Nach einer Verstärkung mit leichter Preemphasis wird das Sprachsignal, wie im Bild 2 gezeigt ist, durch zwei Analogfilter in einen tieffrequenten und einen hochfrequenten Anteil zerlegt. Die getrennte Weiterverarbeitung beider Frequenzbereiche soll die „Verdeckung" der hoch- durch die niederfrequenten Signalanteile und umgekehrt vermindern. Der Messung der Nulldurchgangsabstände geht eine Umwandlung der beiden gefilterten analogen Signale in Rechteckimpulse voraus. Die Flanken der beiden Pulse lösen über die CTC im angeschlossenen Mikrorechner Interrupts aus, deren Abstände mit Hilfe eines CTC-Zählkanals bestimmt werden. Die möglichen Nulldurchgangsabstände in den beiden Frequenzbereichen werden in je vier Intervallklassen (im Bild 2 als Kanäle bezeichnet) eingeteilt, deren Grenzen in der Frequenzebene denen der Frequenzgruppen im Wahrnehmungsbereich des menschlichen Innenohres entsprechen. Das Analyseprogramm ermittelt die Häufigkeit der Repräsentanten jeder Intervallklasse in einem Zeitfenster von etwa 26 ms und bestimmt so ein sogenanntes Nulldurchgangshistogramm z. Durch Vergleich mit einem festen Schwellwert z_s wird jeder der acht Häufigkeitswerte im Nulldurchgangshistogramm in nur einem Bit abgebildet (s. Bild 3). Für jede sprachliche Einheit entsteht so eine Folge von 8-bit-Merkmalvektoren, wobei jeder Vektor einen Signalabschnitt von 26 ms beschreibt und die Anzahl der Vektoren von der Sprechdauer abhängt.

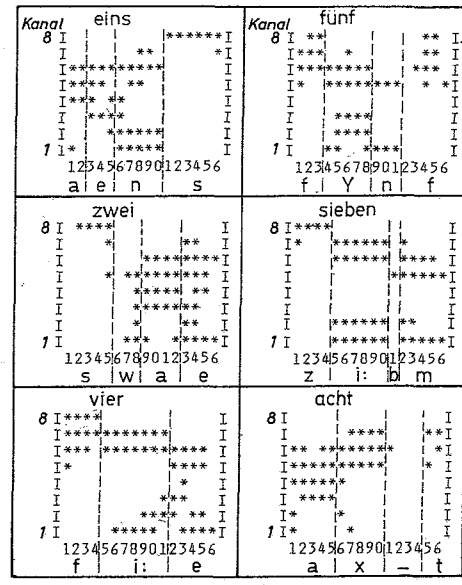

Bild 4: Muster gesprochener Ziffern

Um die variable Signallänge, die aus unterschiedlicher Wortdauer bzw. Sprechgeschwindigkeit resultiert, auszugleichen, schließt sich daran eine Stufe an, die eine lineare Zeitnormierung der Merkmalvektorfolge ausführt (Bild 3). Danach haben die Vektorfolgen aller möglichen Wörter eine einheitliche Länge von 16 Vektoren. Aus jedem Wort wird also ein Muster aus 8 bit × 16 = 128 bit erzeugt.

Bild 4 zeigt reale Muster von gesprochenen Ziffern, in denen Lautgrenzen gekennzeichnet sind.

Lernen

Das Anlernen des Spracherkenners durch den Nutzer selbst dient dem Aufbau eines sprecherspezifischen Referenzdatensatzes, in dem Muster aller geläufigen Aussprachevarianten der Wörter des gewählten Wortschatzes enthalten sein sollten. Der Nutzer spricht dazu die gewünschten Wörter in das Mikrofon und vergleicht das bei der ausgeführten Kontrollerkennung ermittelte Ergebnis mit dem von ihm beabsichtigten. Bei Nichtübereinstimmung ist eine entsprechende Korrektureingabe über eine Tastatur erforderlich. Durch dieses Prinzip bekommt der Sprecher schnell einen Eindruck von der Güte des aktuellen Referenzmustersatzes und der Nichteignung bestimmter Wörter. Er kann auch besonders schwierige Wörter häufiger anlernen. Treten keine Verwechslungen oder Rückweisungen mehr auf, kann er den Lernvorgang abbrechen. Je nach den phonetischen Abständen der gewählten Wörter und der Stabilität der Artikulation wird der Lernprozeß nach drei- bis achtmaligem Sprechen jedes vorkommenden Wortes beendet sein.

Die Durchführung dieses Lernvorganges erfordert mindestens eine numerische Anzeige für die erkannte Klassennummer und eine numerische Tastatur zur Eingabe der korrekten Wortklassennummer im verwendeten Mikrorechner. Für den erarbeiteten Referenzdatensatz sollte eine Speichermöglichkeit vorhanden sein, damit der Lernvorgang nicht bei jedem Einschalten des Erkenners oder Sprecherwechsel wiederholt werden muß. Nach dem Laden eines solchen Referenzdatensatzes ist aber in jedem Fall ein Nachlernvorgang anzuwählen, um die Referenzmuster wieder an die aktuelle Aussprachesituation des Sprechers anzupassen.

Der Lernalgorithmus besteht aus einem gesteuerten Speichern der durch das Sprechen der Lernprobe erzeugten Muster mit ihrer Wortklassenzuordnung im Referenzdatensatz. Es hat sich gezeigt, daß das Speichern eines mittleren Musters je Klasse, auch bei Verwendung einer größeren Informationsmenge je Muster, bedingt durch die lineare Zeitnormierung, nicht zum Erfolg führt. Besser geeignet ist ein Verfahren, bei dem für jede deutlich abweichende Aussprachevariante ein neues Muster abgespeichert wird. Ein Clusteralgorithmus prüft bei jedem gesprochenen Wort, ob es durch ein bereits gespeichertes Muster der gleichen Klasse ausreichend repräsentiert wird oder ob das Muster neu in den Referenzspeicher aufzunehmen ist. Außerdem wird registriert, wie häufig welche Muster zu einem richtigen Erkennungsergebnis geführt haben. Auf diese Weise können Muster, die selten oder nicht mehr vorkommende Aussprachevarianten abbilden, wieder gestrichen werden. Der Referenzmustersatz paßt sich somit während des Lernvorgangs immer an die aktuelle Sprechweise an. Die Aufnahme neuer Muster in den Speicher und das Streichen unnötiger Muster werden durch Schwellwerte gesteuert. Mit ihnen wird ebenfalls eine obere Schranke für die maximal mögliche Musteranzahl je Klasse festgelegt, so daß ein unkontrolliertes Wachsen des Referenzmustersatzes vermieden wird und die Anpassung an den verfügbaren Speicherraum möglich ist.

Erkennen

Ist der Referenzmusterspeicher durch Lernen oder das Einlesen eines Referenzdatensatzes auf den Sprecher eingestellt, kann der Erkenner für die Spracheingabe eingesetzt werden. Der Erkennungsalgorithmus, der schon beim Lernen für die Kontrollerkennung wirksam wurde, beruht auf einem bitweisen Mustervergleich (Hammingdistanz) zwischen dem Muster des eben gesprochenen Wortes und allen Mustern der Wörter aus dem Referenzspeicher. Dabei wird das Referenzmuster mit der geringsten Distanz zum Eingangsmuster gesucht. Unterschreitet diese minimale Distanz einen vorgegebenen Schwellwert (Rückweisungsschwelle), wird die dem entsprechenden Referenzmuster zugeordnete Klassennummer als Erkennungsergebnis in der vereinbarten Speicherzelle abgelegt, andernfalls erfolgt eine Rückweisung. Das Wort ist dann noch einmal zu sprechen. Der Erkennungsalgorithmus wird von der Wortsignaldetektion aktiviert, die interruptgesteuert im Hintergrund arbeitet und das Vorhandensein von Sprachsignalen im Eingangssignal meldet. Er beginnt sofort bei Beginn einer Pause am sogenannten prognostizierten Wortende mit dem Vergleich der Muster. Erst nach dem Ablauf von 200 ms, die zur sicheren Erkennung des Wortendes notwendig sind, wird das Erkennungsergebnis freigegeben. Sollte sich die Wortendprognose als falsch erwiesen haben

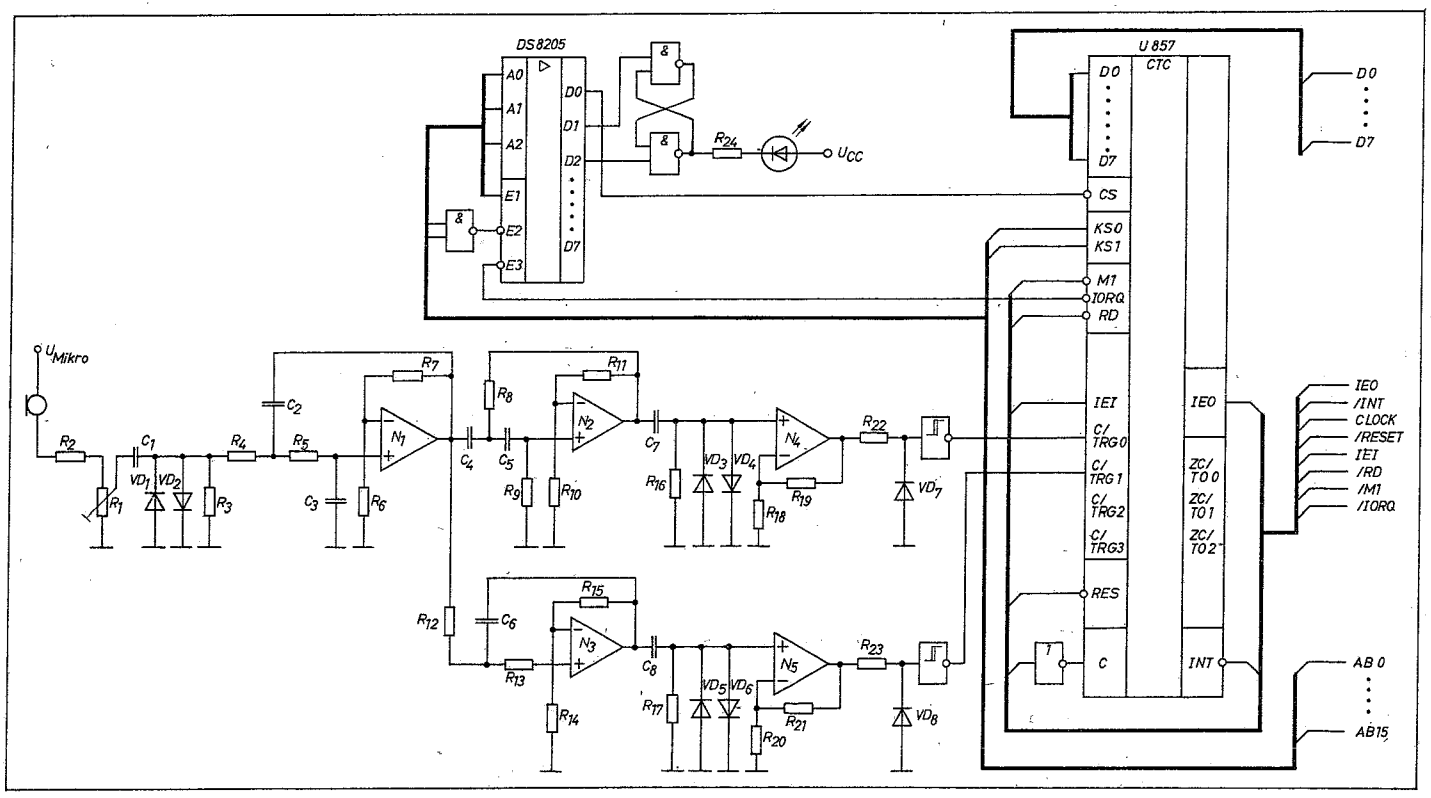

Bild 5: Schaltung des Zusatzmoduls (E 3 und IORQ nicht miteinander verbunden, beide im Steuerbus)

(z. B. nur Verschlußpause vor einem Explosivlaut wie bei „Acht" im Bild 4), beginnt der Erkennungsalgorithmus bei der nächsten Wortendprognose erneut.

Schaltungstechnische Realisierung

Die für den Spracherkenner erforderliche Zusatzschaltung hat folgende Aufgaben zu erfüllen:

- Anpassung des Signalwandlers
- Selektierung des Signals in zwei Frequenzkanäle
- absolute Begrenzung der Signale in beiden Kanälen und Umformung in TTL-Signale
- Unterstützung der Bestimmung von Nulldurchgangshistogrammen
- Anpassung an den Rechnerbus
- optische Bereitschaftsanzeige (kann entfallen).

Das daraus resultierende Schaltbild zeigt Bild 5.

Eingangsverstärker

Der Eingangsverstärker dient der Anpassung des verwendeten Mikrofons. Er besitzt Tiefpaßverhalten, um den Sprachfrequenzbereich auf den verarbeitbaren Bereich zu beschränken. Das hier verwendete Mikrofon SP 75 vom VEB Funkwerk Kölleda besitzt einen eigenen Vorverstärker, so daß der Eingangsverstärker nur aus dem aktiven Tiefpaß besteht, der ab 7 kHz einen Abfall von 12 dB/Oktave aufweist. Die Verstärkung des mit N_1 realisierten Tiefpasses zweiten Grades beträgt 40 dB. Die Resonanzüberhöhung verwirklicht eine leichte Preemphasis, die den Leistungsabfall des Sprachsignals bei hohen Frequenzen mindern soll.

Die Anschaltung der Mikrofonkapsel des SP 75 einschließlich Vorverstärker erfolgt als Zweipol. Mit R_1 kann der Eingangspegel eingestellt werden. VD_1 und VD_2 dienen zu dessen Begrenzung.

Kanaltrennung

Die Aufteilung in die beiden Frequenzkanäle nehmen die durch N_2 und N_3 gebildeten Hoch- bzw. Tiefpässe zweiten Grades vor. Zur Gewährleistung der Stabilität wurde eine Verstärkung von v = 5 festgelegt. Die Trennfrequenz liegt bei 1 kHz, da dadurch die für das Sprachsignal charakteristischen Frequenzen des ersten und zweiten spektralen Maximums in getrennten Kanälen abgebildet werden. Die verwendeten Schaltungen der Filter zeichnen sich durch minimalen Bauelementeaufwand und Unempfindlichkeit gegen Toleranzen der frequenzbestimmenden Bauelemente aus. Zur Dimensionierung wurden die in [1] angegebenen Anwendungskriterien und Entwurfsbedingungen verwendet.

Amplitudenbegrenzung

Die Bildung der für eine digitale Auswertung erforderlichen amplitudenbegrenzten Signale erfolgt mit den Triggerstufen N_4 und N_5. VD_3 bis VD_6 dienen der Pegelbegrenzung. Da die verwendeten Operationsverstärker mit interner Frequenzkompensation eine zu geringe Spannungsanstiegsgeschwindigkeit aufwiesen, wurden zur Versteilerung der Flanken TTL-Schmitt-Trigger nachgeschaltet.

VD_7 und VD_8 begrenzen die Eingangsspannungen auf den TTL-Pegelbereich.

Digitalteil

Der Digitalteil unterstützt einerseits die Bildung der Nulldurchgangshistogrammfolgen, andererseits realisiert er die Ankopplung des Spracheingabemoduls an den Bus des verwendeten U-880-Mikrorechners. Grundelement des Digitalteils ist die Zähler-Zeitgeber-IS U 857 D (CTC). Die Kanäle 0 und 1 empfangen die Flanken der Eingangssignale für die Messung der Nulldurchgangsabstände. Kanal 2 dient als Zeitgeber für die Dauer der Analysezeitfenster. Über die Bussignale der CTC ist ein problemloser Anschluß an den steuernden Rechner möglich. Hier sind Treiber einzuschalten, die im Bild 5 nicht gezeigt sind. Der Spracheingabemodul belegt sechs E-A-Adressen. Neben den vier für die CTC sind noch zwei für das Ein- bzw. Ausschalten einer Sprechaufforderungsanzeige mit VD_9 erforderlich, die über ein RS-Flip-Flop angesteuert wird. Die verwendeten Adressen sind mit Rücksicht auf einen minimalen Bauelementeaufwand nicht wählbar.

Abgleich der Baugruppe

Der Eingangspegelsteller R_1 ist so einzustellen, daß die Hoch- bzw. Tiefpaßstufe auch bei Pegelspitzen mit unbegrenztem Signal angesteuert werden. Die Einstellung der Triggerschwellen von N_4 und N_5 erfolgt unterschiedlich. Um beide Kanäle gleich auszusteuern, haben sich im Hochpaßkanal eine Schwelle von −28 dB bei 2 kHz und im Tiefpaßkanal von −18 dB bei 500 Hz gegenüber Vollaussteuerung als optimal erwiesen. Im praktischen Einsatz ist eine Anpassung der Einstellung von R_1 an das Umgebungsgeräusch empfehlenswert.

Programm und Rechnereinbindung

Programmstruktur

Das Programm für den Betrieb des Spracherkenners umfaßt ein Hauptprogramm für das Anlernen mit je einem Eingang für das Neulernen (leerer Referenzspeicher) und das Weiterlernen (aufbauend auf schon vorhandenem Referenzwissen) und ein Unterprogramm für die Erkennung eines Signalabschnittes. Es ist als Assemblerquellprogramm verfügbar. Das zugehörige Maschinenprogramm mit einem Speicherbedarf von 2 Kbyte ist PROM-fähig und kann auf beliebige 1-Kbyte-Grenzen im Speicherraum des Wirtsrechners geladen werden. Es wird ein Arbeitsspeicher von 4 Kbyte benötigt, der an beliebigen 1-Kbyte-Grenzen beginnen kann. Das Programm gliedert sich in drei wesentliche funktionelle Segmente:

- Anpassung an Wirtsrechner
- Steuerroutinen für Lernen und Erkennen
- Unterprogramme für Signalanalyse, Erkennung und Lernen.

Die ersten beiden Segmente können durch den Nutzer modifiziert werden.

Programmschnittstellen

Die Eingänge des Hauptprogramms LERNEN werden über folgende Adressen aufgerufen (mit JMP):

- NEULERN: Lernen mit anfangs leerem Referenzspeicher
- WEILERN: Lernen, aufbauend auf vorhandenem Referenzwissen.

Nach Abschluß des Lernprogramms erfolgt ein Sprung zu einer vereinbarten Adresse (z. B. Warmstart des Betriebssystems). In den 4 Kbyte des Arbeitsspeichers liegt ein Referenzdatensatz vor, der auf ein peripheres Medium ausgelagert werden bzw. als Basis für Abarbeitung des Unterprogramms RECOG (Erkennung eines Signalabschnittes) dienen kann.
Eine Parameterübergabe ist bei dessen Aufruf nicht notwendig. Mit dem Aufruf wird die Eingabebereitschaft des Signalanalysators hergestellt. RECOG wird verlassen nach Abschluß der Erkennung (spätestens 200 ms nach Überlauf des Eingabepuffers für 1,8 s Signaldauer oder nach Erkennung der Pause am Ende eines Signalabschnittes), wenn nicht auf Rückweisung erkannt wird. Eine Rückweisung kann ein akustisches Signal auslösen und führt zur erneuten Eingabebereitschaft. Nach der Rückkehr aus dem Unterprogramm liegt im A-Register und in einer vereinbarten Speicherzelle das Symbol der erkannten Signalklasse vor (Übergabe auch für Basic geeignet).

Einbindung in den Wirtsrechner

Die unmittelbare Anpassung des Spracherkennerprogramms an die Umgebung im Wirtsrechner erfolgt durch Modifikation des Assemblerquellprogramms an deutlich gekennzeichneten Stellen. Es sind z. B. folgende Speicheradressen anzugeben:

PROM: Beginn des Programmbereiches (2 Kbyte)
RAMA: Beginn des Arbeitsspeichers (4 Kbyte)
OPSYS: Rückkehradresse aus dem Lernen
ERKER: Adresse der Speicherzelle für die Übergabe des Erkennungsergebnisses (1 byte)
CTKO: Beginn des Freibereiches in der Interrupttabelle des Betriebssystems (6 byte für CTC-Kanäle 0 bis 32).

Drei aufeinanderfolgende Adressen im E-A-Adreßraum sind zu suchen:

CTCO: erste der drei E-A-Adressen, die auch im Adreßdekoder des Moduls einzustellen ist.

Folgende Kommunikationsunterprogramme sind mit Hilfe des Betriebssystems des Wirtsrechners zu realisieren:

TEIBD: Eingabe eines Zeichens von der Tastatur
TAUBD: Ausgabe eines Zeichens zur Anzeigeeinrichtung
BEEP: Erzeugung eines akustischen Signals
BEREI: Einschalten einer Bereitschaftsanzeige und Blockierung aller Interruptquellen im Wirtsrechner
BERAU: Ausschalten der Bereitschaftsanzeige und Freigabe der mit BEREI gesperrten Interruptquellen
INKEY: Abfrage Tastaturstatus für den Abbruch einer laufenden Signaleingabe.

Neben diesen programmtechnischen Anpassungen können zwei interne Programmkonstanten durch den Nutzer experimentell variiert werden, um eine optimale Anpassung des Erkenners an die Einsatzbedingungen zu erreichen:

RUESW: Rückweisungsschwelle
SWCLU: Schwellwert für das Streichen von Mustern aus dem Referenzsatz. Eine Vergrößerung des Standardwertes verringert die Musteranzahl, mit der eine Klasse im Referenzspeicher vertreten ist. Der notwendige Speicherbedarf für die Speicherung eines Klassenvorrates wird so gemindert, gleichzeitig wird jedoch die mögliche Variationsbreite innerhalb einer Singalklasse eingeschränkt.

Neben dieser Anpassung an die Rechnerumgebung sollte vor allem der Aufruf des Unterprogramms zur Erkennung eines Wortes optimal in die vom Anwendungsfall abhängige Kommunikation eingebunden werden, um die Effekte der Spracheingabe voll zu nutzen.
Je nach Anwendungsfall kann der Nutzer z. B. den Spracherkenner auf verschiedene Arten aktivieren. Der Erkenner ist entweder immer aktiv, oder er wird erst durch eine zusätzliche Steuerinformation (z. B. Taste oder Fußschalter) oder ein Schlüsselwort freigegeben. Besondes bei impulshaltigem Lärm oder anderen notwendigen sprachlichen Äußerungen des Nutzers ist es nicht günstig, wenn der Erkenner permanent aktiv ist. Ein ungewolltes Ansprechen ist dann nicht auszuschließen. Für solche Fälle hat sich die Aktivierung durch ein Schlüsselwort besonders bewährt. In dieser Betriebsart ist der Spracherkenner erst dann bereit, Erkennungsergebnisse abzugeben, wenn er aus der Fülle der auf ihn einwirkenden Geräusche ein bestimmtes Schlüsselwort erkannt hat. Dieses Schlüsselwort kann der Nutzer selbst wählen und muß es mit anlernen. Bei geschickter Wahl des Schlüssels (stark strukturiertes Wort) ist die Wahrscheinlichkeit eines ungewollten Ansprechens sehr gering. Da die Eingabegeschwindigkeit sich bei dieser Betriebsart verringert, eignet sie sich vor allem dort, wo verhältnismäßig wenig Eingaben je Zeiteinheit nötig sind. Die Eingabegeschwindigkeit kann erhöht werden, wenn der Spracherkenner durch ein Schlüsselwort freigegeben wird und dann aktiv bleibt, bis ein Abschlußwort erkannt wird. Innerhalb der durch Schlüssel- und Abschlußwort gebildeten Zeitspanne können beliebig viele Wörter eingegeben werden.

Besonderheiten der Bedienung

Beide Lernprogramme zeigen nach dem Start die Anzahl der im Arbeitsspeicher abgelegten Referenzmuster an (max. 200). Dem Nutzer wird damit eine Information über weitere Möglichkeiten zum Nachlernen gegeben. Durch einen Abbruch des Lernvorgangs und einen folgenden Aufruf von WEILERN ist eine ständige Kontrolle des Füllstandes des Referenzspeichers möglich.
Anschließend wird der Nutzer zur Eingabe der Probenkennung aufgefordert. Die Probenkennungen 0 bis 4 sind möglich. Ein Referenzsatz kann also aus maximal fünf unabhängigen Lernproben aufgebaut werden. Die Kennzeichnung der Probe dient der notwendigen Separierung der von verschiedenen Nachlernvorgängen stammenden Referenzmuster, um Verdeckungseffekte zu vermeiden (Streichung von noch wertvollen Mustern aus anderen Aussprachesituationen).
Durch die immer sicherer werdende Erkennung mit zunehmender Zahl der Lernproben ist der Erfolg des Lernvorgangs zu beobachten. Ebenso wird deutlich, welche Klassen sich durch den Erkenner nicht unterscheiden lassen. Dies sind phonetisch ähnliche Wörter, von denen dann eines durch ein Synonym ersetzt werden muß (z. B. „zwo" statt „zwei").
Zur Erreichung bestmöglicher Erkennungsergebnisse sollte der Lernvorgang erst abgebrochen werden, wenn alle Realisierungen der Lernstichprobe stabil richtig wiedererkannt werden. Ein objektives Maß für die Beendbarkeit des Lernens ist jedoch nur das Abnehmen der Musteranzahl bei weiterer Eingabe von Lernproben, d. h., der Erkenner „weiß" genug über die Signalquelle. Dies ist über die Anzeige der Musteranzahl beim Start von WEILERN kontrollierbar. Konvergiert die Musteranzahl nicht, artikuliert der Sprecher zu ungleichmäßig, oder die Signale sind durch Fremdeinflüsse zu stark gestört. Der Lernvorgang muß dann bei gefülltem Referenzspeicher abgebrochen werden.
Von wesentlicher Bedeutung für die Erkennungssicherheit bei der Verarbeitung von Lautsprache ist die Auswahl eines geeigneten Mikrofons und die Arbeit mit diesem. Es ist unbedingt ein mundnah zu tragendes, nahbesprechbares Mikrofon (z. B. SP 75 vom VEB Funkwerk Kölleda) zu verwenden. Das Mikrofon darf seine Lage (etwa 1 cm seitlich vor dem Mund) durch Bewegung des Kopfes nicht verändern und muß diese Lage auch nach dem erneuten Aufsetzen wieder einnehmen.

Einsatzergebnisse

Der Spracherkenner-Zusatzmodul wurde als billige, nachnutzbare Variante des Einplatinen-Spracherkenners ESE K 7824 von VEB Robotron Elektronik Dresden entwickelt. Mit einem international üblichen Test nach [2] wurde eine Erkennungsquote von 97% erreicht. Den ESE K 7824 erprobten Anwender in verschiedenen Bereichen der Volkswirtschaft. Die Erprobungsergebnisse sind in [3] ausführlich dargestellt.
Der Spracherkenner-Zusatzmodul wurde vom VEB Robotron-Meßelektronik „Otto Schön" Dresden für einen Spracheingabemodul des Kleincomputers KC 87 nachgenutzt, der noch 1987 angeboten werden soll.

Literatur

[1] Herpy, M.: Analoge integrierte Schaltungen. Budapest: Akadémiai Kiadó, 1976
[2] Doddington, G. R.; Schalk, T. B.: Speech recognition: turning theory to practice. IEEE spectrum, Philadelphia 18 (1981) 9, S. 26–32
[3] Seveke, L.: Einsatz von Spracherkennern. Nachrichtentechnik – Elektronik, Berlin 37 (1987) 1, S. 34–36