EIN EXPERIMENTIERSYSTEM ZUR SPRACHSIGNALANALYSE UND -ERKENNUNG

St. Kürbis; V. Paul
Technische Universität Dresden

Mit wachsenden Aktivitäten auf dem Gebiet der Sprachsignalverarbeitung und -erkennung stiegen Umfang und Komplexität der zur Verfügung stehenden Verarbeitungsmöglichkeiten. Dies ist auf mehrere Gründe zurückzuführen: Zum einen erlaubt die sehr schnell wachsende Leistungsfähigkeit der modernen Rechentechnik die Implementierung nahezu beliebig komplizierter Algorithmen, deren praktische Nutzung noch vor wenigen Jahren nicht möglich gewesen wäre. Dazu zählen zum Beispiel die Anwendung digitaler Signalanalyseverfahren ebenso wie die Implementierung von Fuzzy-Algorithmen oder die komplexe Modellierung neuronaler Verarbeitungsstrukturen. Zum anderen wurde der Fundus an theoretischen Ansätzen für die Signalverarbeitung ständig weiterentwickelt und erweitert. Dazu gehören sowohl formale und formal statistische Algorithmen wie die dynamische Optimierung ,die Hidden-Markov-Modellierung oder neuronale Netzwerkverfahren als auch Verfahren, die sich stark an die humane Sprachverarbeitungsvorgänge anlehnen und Erkenntnisse psychoakustischer Untersuchungen ausnutzen.

Ergebnis dieser Entwicklung ist eine beachtliche Menge von Teilerkenntnissen und entsprechenden, daraus abgeleiteten Algorithmen (teilweise auch erst im Ansatz) für einzelne Stufen der Verarbeitung und Erkennung akustischer Sprachsignale. Daraus ergeben sich für wissenschaftliche Grundlagenuntersuchungen zur Entwicklung eines leistungsfähigen Spracherkennungssystems zwei aufeinander aufbauende Grundforderungen:

1. Es muß eine generelle Grobarchitektur, eine "Systemphilosophie" erarbeitet werden, die im Interesse breit angelegter Untersuchungen möglichst viele verschiedene Konkretisierungen zuläßt.

2. Der praktische Rahmen für die Implementierung von zu erprobenden Verfahren sollte aus dem o.g. Grund in Form eines Experimentiersystems erfolgen, das ein möglichst problemfreies Austauschen bzw. Ergänzen von neuen Algorithmen erlaubt und auf dessen Grundlage dann z.B. sowohl Einzelwort- als auch Wortfolgenerkenner aufgebaut werden können.

Die Grundidee und die realisierte Implementierung des an der TU Dresden in der Forschungsgruppe "Sprachkommunikation" entstandenen Experimentiersystems zur Sprachsignalerkennung soll im folgenden kurz dargestellt werden.

1. Zur Systemphilosophie

Grundanliegen des Systems ist die Erkennung gesprochener Sprache (im Gegensatz zur Schriftsprache), d.h., die Zuordnung eines akustischen Signals zu einer symbolischen Beschreibung (z.B. Klassennummern oder Folgen von Klassennummern). Im Bild 1 ist die Grobstruktur des verwendeten Systemkonzepts dargestellt. Dem System zugrunde gelegt werden sogenannte Basiselemente. Unter Basiselementen sollen hier sprachliche Fragmente wie Halbsilben, Silben, Wortfragmente oder Worte verstanden werden.

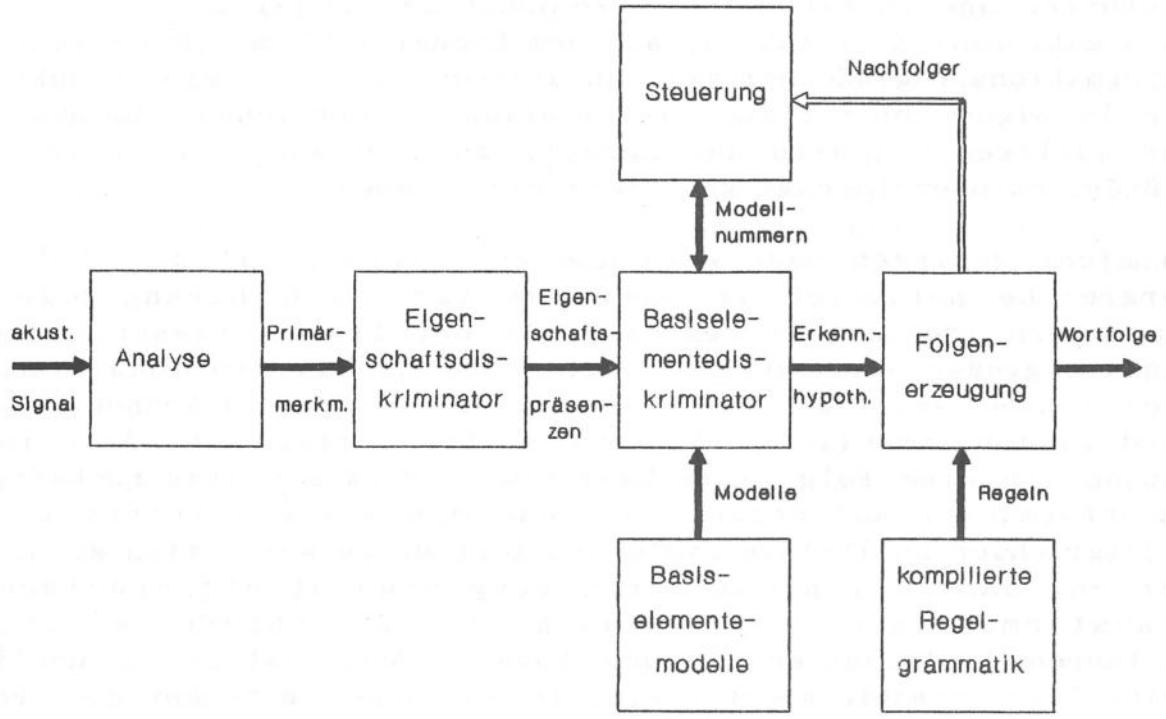

Bild 1: Grobstruktur des Experimentiersystems

Kernidee der verfolgten Strategie ist dabei der Vergleich von aus dem akustischen Signal gewonnenen Parametern (Merkmalvektorfolge) mit im System gespeicherten Modellen von Basiselementen. Dabei kann es zur Abbildung einer (vollständigen) Merkmalvektorfolge auf einzelne (Wort-) Modelle kommen (Einzelworterkennung) oder aber die Eingangsfolge wird auf mehrere aufeinanderfolgende Modelle abgebildet (Wortfolgenerkennung oder Worterkennung auf der Basis kleinerer Einheiten). Kernstück ist dabei ein Vergleichsblock, der im allgemeinsten Fall Teile der ihm angebotenen Eingangsfolge mit Modellen von Basiselementen vergleicht und daraus Erkennungshypothesen ableitet. Mit Hilfe einer übergeordneten Steuerung wird es möglich, die gespeicherten

Basiselementemodelle zu Modellen größerer Einheiten zu verketten. Diese Verkettung kann ungesteuert erfolgen, d.h., es ist jede beliebige Folge von Basiselementen möglich. Das ist z.B. bei der Verkettung von Wortmodellen zu Sätzen denkbar. Es ist aber auch möglich, diese Verkettung zielgerichtet zu steuern, was sowohl bei der Verknüpfung von Wortfragmenten (Silben) oder kleineren Einheiten zu Worten sinnvoll ist (z.B. anhand eines entsprechenden Wörterbuches) als auch die Erkennung von zu Sätzen verbundenen Worten erheblich verbessern bzw. vereinfachen kann. Grundlage für die Steuerung bildet eine in kompilierter Form gespeicherte Regelgrammatik mit Verknüpfungsvorschriften.

Ausgangspunkt aller Verarbeitungen ist das akustische Signal. Dieses ist Träger der Information über den Inhalt der sprachlichen Äußerung, enthält gleichzeitig aber auch Redundanz, Störungen und im Sinne der Erkennung des Inhalts irrelevante Informationen. Ziel ist es, aus dem Signal im Sinne der Erkennung informationstragende Merkmale abzuleiten und damit eine Reduktion der im Signal enthaltenen Information zu erreichen. Am Ausgang der Analysestufe wird das Sprachsignal durch eine Folge von Meßwertvektoren (Merkmalvektoren) beschrieben.

In nicht geringem Maße wird die Zuverlässigkeit der Modellvergleiche natürlich von den zum Vergleich herangezogenen Parametern (Merkmalen) bestimmt. In der Systemkonzeption wurde deshalb großer Wert auf die Möglichkeit gelegt, vor dem eigentlichen Modellvergleich Transformationen einfügen zu können. Diese wird in der darauffolgenden Stufe, der Eigenschaftsdiskrimination, in eine Folge sprachangepaßter Merkmale (Eigenschaften) transformiert, auf deren Grundlage die weitere Verarbeitung erfolgt. Hier sollte die Systemarchitektur so weit offen gehalten werden, daß ggf. nicht nur streng sequentiell arbeitende Transformationen, sondern auch in Zeitrichtung selektive Verfahren implementiert werden können. Auch sollte es möglich sein, Transformationsmodule einzeln oder aber in verketteter Form (mehrere Transformationen nacheinander) einzubinden.

2. Komponenten des Experimentiersystems

2.1. Analysestufe

Am Eingang der Analysestufe liegt das Sprachsignal in analoger Form als Spannungs- Zeit- Funktion vor. Für die weitere Verarbeitung kann eine obere Grenzfrequenz von 8 kHz und eine Dynamik von mindestens 60 dB angenommen werden, was einen Datenfluß von etwa 200.000 bit je Sekunde am Eingang des Analysators bedingt.

Dieser Datenstrom wird in der Analyse meist in äquidistante Zeitabschnitte ("Zeitfenster") zerlegt, wobei die Länge dieser Zeitabschnitte sowie der Abstand aufeinanderfolgender Abschnitte wählbare Analyseparameter darstellen. Jeder dieser Abschnitte

wird durch einen Satz von Meßwerten beschrieben. Verschiedene Faktoren bedingen den Einsatz unterschiedlicher Verfahren zur Gewinnung der Meßwerte:

1. Die Art der verwendeten Basiselemente (Laute, Silben, Worte,...), deren Unterscheidbarkeit und daraus folgend die im Eigenschaftsdiskriminator zu beschreibenden (sprachspezifischen) Merkmale erfordern signalspezifische Verfahren.

2. Der Kenntnisstand über die Verschlüsselung der erkennungsrelevanten Informationen bestimmt wesentlich die erforderliche Meßwertauswahl.

3. Echtzeitverhalten der implementierten Analyseverfahren, erforderliche Gerätetechnik sowie damit verbundene Kosten unterscheiden sich stark bei den unterschiedlichen Verfahren.

Die Analysekomponente des Experimentiersystems unterstützt die Implementierung mehrerer beliebiger Analyseverfahren. Ausgehend von der digital zwischengespeicherten Zeitfunktion können vom Systemnutzer Verfahren wie Spektralanalyse (FFT, Filterbank), Korrelationsanalyse, LPC- Analyse und Cepstrum- Analyse auf das Sprachsignal angewendet werden. Durch die Bereitstellung geeigneter Schnittstellen können vom Nutzer beliebige Algorithmen kombiniert bzw. ergänzt werden. Der hohen Flexibilität bei dieser Vorgehensweise steht jedoch eine hohe Analysezeit gegenüber. Für die Untersuchung umfangreicher Sprachproben wurden Analysemoduln geschaffen, die wesentliche Analyseverfahren mit Hardwareunterstützung in Echtzeitnähe realisieren. Dazu gehören z.Zt. Parallelfilteranalyse mit verschiedenen Zeit- und Frequenzauflösungen, FFT-Analyse einschließlich Vor- und Nachverarbeitung, Cepstrumanalyse und Grundfrequenzanalyse nach dem AMDF-Verfahren. Diese Analysemoduln sind miteinander koppelbar, so daß auch kombinierte Merkmalvektoren zur Signalbeschreibung erzeugt werden können.

2.2. Eigenschaftsdiskrimination

Die Aufgabe dieser Stufe soll darin bestehen, aus den von der Analyse gelieferten Meßdatenfolgen solche Merkmale zu gewinnen, die für die Wiedererkennung der sprachlichen Äußerung geeignet sind. Die Notwendigkeit dieses Schrittes ergibt sich aus der mangelnden Reproduzierbarkeit der physikalischen Eigenschaften durch den Menschen als Signalquelle. Das Problem dabei ergibt sich insbesondere aus der Tatsache, daß bei einer hochgenauen Ermittlung der physikalischen Signalparameter zwar alle informatonstragenden Eigenschaften (mit) abgebildet werden, zusätzlich aber auch eine Vielzahl von sprecherspezifischen, situationsspezifischen und umgebungsspezifischen Informationen mit erfaßt werden, die für die eigentliche Spracherkennung als Störgrößen aufgefaßt werden müssen. Diese Nebeninformationen

können zwar teilweise durch eine Verringerung der Analysegenauigkeit unterdrückt werden (z.B. geringere Zeit-, Frequenz-, Amplitudenauflösung), gleichzeitig werden damit aber auch erkennungsrelevante Informationen unterdrückt. Mit Hilfe einfacher Vorverarbeitungsverfahren lassen sich hier einige Kompromisse finden. Dazu gehören z.B. Algorithmen zum Ausgleich von Lautstärkeschwankungen und einfache Filterstrukturen zur Dämpfung impulsartiger Störungen (Knackgeräusche...). Mit diesen relativ unkomplizierten Verfahren können Verbesserungen der Systemeigenschaften bis zu einer gewissen Stufe erreicht werden. Wesentlich komplizierter und derzeit auch erst im Anfangsstadium sind solche Verfahren, die auf der Grundlage angesammelten Wissens über die charakteristischen Eigenschaften menschlicher Sprache eine komplexere Auswertung der Analysedaten vornehmen. Dazu können die Ergebnisse langjähriger Arbeiten am Sprachsignal zugrunde gelegt werden. Im Unterschied zu den einfacheren Transformationskriterien könnten hierbei in wesentlich stärkerem Maße zeitliche Parameterverläufe berücksichtigt werden. Als Beispiel sollen hier Überlegungen angeführt werden, wie die Ähnlichkeit zu einem stimmlosen Plosiv als eine sprachspezifische Eigenschaft ermittelt werden könnte. Derartige Laute sind ja bekanntlich dadurch gekennzeichnet, daß nach einer 30...100 ms langen energiearmen Phase ("Verschlußpause") ein sehr starker Anstieg der Energie beobachtet werden kann, bei dem der Frequenzbereich oberhalb 2 kHz dominiert. Nach ca. 30...50 ms sinkt die Energie dann wieder auf ein Niveau ab, das vom nachfolgenden Laut bestimmt wird. Diese aus der Sicht der künstlichen Intelligenz als prozedurales Wissen ansehbaren Kenntnisse können in einen geeigneten Algorithmus umgesetzt werden, dessen Ausgabeinformation eine Eigenschaftspräsenz "stimmloser Plosiv" wäre. In dieser oder ähnlicher Weise sind Prozeduren vorstellbar, die die Zugehörigkeit des analysierten Signalabschnittes zu bestimmten Eigenschaftsklassen bestimmen. An dieser Stelle sei auch verwiesen auf die Arbeiten von LANGMANN /1/ zur wissensbasierten Untersuchung von Sprachsignalen, bei der solche Eigenschaftspräsenzen mit Hilfe von Fuzzy-Methoden in einem sehr leistungsfähigen Expertsystem ermittelt werden. Vorstellbar ist auch die Implementierung von Verfahren, die zunächst aus einer Auswahl aus der angebotenen Meßdatenmenge Eigenschaften ableiten, bei Unsicherheiten bzw. für die bestbewertetsten Eigenschaften in einem zweiten Schritt mit Hilfe zusätzlich herangezogener Primärdaten eine Verifikation bzw. Präzisierung dieser Eigenschaften vornehmen.

Der derzeitige Realiserungsstand der Eigenschaftsdiskrimination erstreckt sich auf formale Vorverarbeitungsroutinen (Pegelausgleich, Glättung, Störunterdrückung) als nutzbare Module sowie die experimentellen Arbeiten von LANGMANN /1/ zur wissensbasierten Eigenschaftsselektion. Alle weiteren Vorstellung tragen konzeptionellen Charakter und bilden derzeit einen wesentlichen Schwerpunkt der Arbeiten der Forschungsgruppe.

2.3. Klassifikation

Der einfachste Fall einer Klassifikation ist die Einzelworterkennung. Dabei wird eine Vektorfolge begrenzter Länge vollständig auf einen Vorrat von Vergleichsfolgen (Referenzmodellen) abgebildet und ein Maß für die Übereinstimmung (oder Abweichung) jedes Modells ermittelt. Daraus kann eine Erkennungshypothese nach dem Kriterium minimaler Abweichung bzw. maximaler Ähnlichkeit abgeleitet werden. Bereits für den Fall dieser Einzelworterkennung ist die Auswahl der Realisierungsmöglichkeiten beträchtlich. Erwähnt werden sollen hier die unterschiedlichen Ansätze zur Bestimmung von Ähnlichkeits- und Distanzmaßen, unterschiedliche Modellansätze (Matrizen, Markov-Ketten, Hidden-Markov-Modelle u.a.m.) sowie unterschiedliche Abbildungsverfahren (DP, Matrixvergleich...).

Erheblich aufwendiger wird diese Struktur bei der Verarbeitung von Modellketten. Es soll zunächst offen bleiben, ob es sich dabei um die Verkettung von Wort- zu Satzmodellen oder um die Verkettung von Modellen kleinerer Basiselemente zu Worten handelt. In beiden Fällen ist für die zu verarbeitende Sprachäußerung kein geschlossenes Modell vorhanden. Dieses muß in irgendeiner Form aus den Modellen der Basiselemente aufgebaut werden. Voraussetzung dafür ist ein entsprechender Aufbau dieser Elementarmodelle, der ein Verketten gestattet. Sofern die Modelle die Form implizit oder explizit gespeicherter Parameter-Zeit-Verläufe besitzen, ist ein solches Verketten in der Regel ohne Probleme möglich. Als weitere Voraussetzung muß angesehen werden, daß der eigentliche Vergleichsalgorithmus Modellketten theoretisch beliebiger Länge verarbeiten kann, also sequentiell arbeitet.

Ein wichtiges Kriterium ist die Frage, ob die Verkettung von Basiselementemodellen zu Modellketten ungesteuert erfolgt, also z.B. Worte in beliebiger Reihenfolge zu Sätzen oder Silben beliebig zu Worten aneinandergereiht werden können. Als Alternative steht hier die Einbeziehung von linguistischem Wissen in den Prozeß der Modellfolgenerzeugung. Das Systemkonzept sieht hier folgende Strategie vor:
Der Folgenvergleich beginnt mit der Bereitstellung der zu betrachtenden Elementarmodelle durch eine Steuerung (vgl. Bild 1). Mit diesen angegebenen Basiselementemodellen wird die Eingangsfolge Vektor für Vektor verglichen. Dabei wird irgendwann der Zeitpunkt erreicht, daß bei optimaler Abbildung der verarbeiteten Vektor-(Teil-)folge das Ende eines oder mehrerer Modelle erreicht wird. Darüber wird die Steuerung informiert, die daraufhin dem Vergleichsalgorithmus die möglichen anzuschließenden Modelle mitteilt. Außerdem wird das beste zu jedem Zeitpunkt beendete Modell als Erkennungshypothese für den verarbeiteten Signalabschnitt ausgegeben. Aus diesen Erkennungshypothesen wird durch den abschließenden Verknüpfungsalgorithmus die Ergebnisfolge aufgebaut. Es sind zwei Grundvarianten dieses Komplexes möglich:

a) Die Verkettung der Elementarmodelle erfolgt ohne Berücksichtigung der jeweiligen Vorgängermodelle. In diesem Falle entfällt die (im Bild 1 als nicht ausgefüllter Pfeil dargestellte) Rückkopplung von der Folgeerzeugungsstufe zur Steuerung. Diese erzeugt selbständig alle durch Kombination der Elementarmodelle möglichen Modellketten. In diesem Falle fällt der Folgenerzeugungsstufe die Rolle einer "Nachverarbeitung" zu, die die angebotenen Erkennungshypothesen- Folgen auf Regelkonsistenz zu überprüfen bzw. regelgerechte Folgen daraus auszuwählen hat.

b) Die im Modul "Folgenerzeugung" zunächst auch erst als Hypothesen aufgebauten Ergebnisfolgen werden analysiert und daraus alle diejenigen Basiselemente ausgewählt, die eine regelgerechte Fortsetzung dieser hypothetischen Folge darstellen können. Diese Informationen werden dem Steuerteil übermittelt und damit dessen Freiheitsgrade bei der Modellverkettung drastisch eingeschränkt. In diesem Falle bilden die drei Komplexe Basiselementediskriminator, Folgenerzeugung und Steuerung eine Einheit, die als "Klassifikator mit regelgesteuerter Modellfolgenerzeugung" charakterisiert werden kann.

3. Implementierung, technische Basis

Das System wurde realisiert als Softwarepaket auf Rechnern vom Typ Robotron K1630 (entspricht etwa PDP 11). Dabei wurden die einzelnen Systemmodule als selbständig lauffähige Tasks organisiert, die über Intertask-Kommunikation Daten austauschen. Auf diese Weise wurde die Konfigurierung einer ausgewählten Modulkombiation zur Laufzeit möglich.

Die Primäranalyse ist hardwaregestützt und basiert zum großen Teil auf Eigenbau-Baugruppen, die in ein 8-bit Mikrorechnersystem integriert wurden. Mehrere dieser Mikrorechner vom Typ K1520 (Basis Z 80 der Firma ZILOG) bilden einen Analysekomplex, der die gleichzeitige Ausführung mehrerer Analyseverfahren gestattet.

Zur Darstellung von Analysedaten, aber auch von inneren Abläufen im Klassifikator werden ebenfalls selbstentwickelte Farbgrafik-Module (auch auf Basis Z 80) mit 256 x 256 Punkten Bildauflösung bei 16 Farben (aus 4096 Farben wählbar) verwendet.

4. Zum gegenwärtigen Stand

Für jeden der unter Punkt 2 beschriebenen Komplexe wurden mehrere verschiedene konkrete Algorithmen als Software- Module realisiert. Die zur Zeit bestehende Modulbibliothek umfaßt u.a. Analysemodule für Parallelfilter-, FFT-, Akustikprozessor und Nulldurchgangsanalyse. Für die Vorverarbeitung wurden z.B. Module zur linearen und nichtlinearen Dynamikanpassung, zur Glättung der

verarbeiteten Daten, zur Kontrastverschärfung und zur Reduktion der zeitlichen Informationsdichte implementiert. Als Klassifikationsmodule wurden drei Grundvarianten realisiert: Ein Einzelworterkenner, der gleichzeitig als Lernprogramm zur Erzeugung numerischer Referenzinformation verwendet werden kann, ein Wortfolgenerkenner, der nach dem Prinzip der wissensgesteuerten Referenzfolgenerzeugung arbeitet und die so generierten Folgen mittels dynamischer Optimierung mit der Eingangsfolge vergleicht und schließlich ein Wortfolgenerkenner mit freier Modellverkettung, der als Ausgangsinformation je Realisierung eine Bewertungsmatrix liefert, die dann mit Hilfe eines syntaktisch gesteuerten Auswerteprogramms in eine Ergebnisfolge überführt wird.
Mit diesen Modulen wurden eine Reihe konkreter Konfigurationen realisiert und experimentell getestet. Die dabei erreichten Erkennungsergebnisse liegen für die Einzelworterkennung bei ca. 90%, bei der Wortfolgenerkennung mit syntaktischer Steuerung bei ca. 60% und bei der Wortfolgenerkennung mit syntaktischer Nachverarbeitung bei ca. 65%.

Literatur

/1/ LANGMANN,D.: Sprachsignalauswertung mit dem Expertsystem EXA 1600; im vorliegenden Band