DIE BEURTEILUNG DER LEISTUNGSFÄHIGKEIT VON SPRACHERKENNUNGS-SYSTEMEN

Helmut Mangold
Daimler-Benz, Forschungsinstitut Ulm

1. Leistungsmessungen bei technischen Systemen

Für den Hersteller wie den Benutzer eines technischen Systems ist es gleichermaßen wichtig, die Leistungsfähigkeit eines technischen Systems möglichst zuverlässig beurteilen zu können. Der Hersteller möchte seinem Kunden gewisse Leistungsparameter verbindlich zusichern und der Kunde möchte wissen, ob die von ihm beabsichtigte Anwendung mit dem geplanten System auch realisierbar ist. Wir benötigen also verbindliche Maßsysteme und ebenso verbindliche Meßvorschriften.

Bei Spracherkennungssystemen gibt es hier noch ganz beachtliche Probleme. In technischen Unterlagen heute erhältlicher Systeme wird zwar meist eine Erkennungsrate zwischen 95 und 100% richtig erkannter Wörter angegeben, doch wird über die Art des verwendeten Testvokabulars, die beim Test verwendeten Randbedingungen und Sprechweisen, die Zahl der Testsprecher oder gar über die Verteilung der Erkennungsleistung unter den verwendeten Testsprechern praktisch nur selten etwas gesagt.

In /1/ werden etwa 80 Faktoren zusammengestellt, die auf die Erkennungsleistung eines Spracherkenners Einfluß haben. Auf der anderen Seite können wir bei der Durchsicht eines neueren Tagungsbandes /2/ eine Fülle von Erkennungsraten für die verschiedensten Spracherkennungssysteme und die unterschiedlichsten Randbedingungen finden. Wir finden da Werte von fast 100 % für einfache Randbedingungen aber auch von annähernd 0 %, wenn etwa einzeln gesprochene Wörter in stark gestörter Umgebung erkannt werden sollen. Da gibt es Systeme, die an den zehn gesprochenen Ziffern getestet werden, bei anderen Systemen wurde ein gewaltiges Vokabular von einigen tausend Wörtern verwendet. Hier wird es dann nicht nur für einen potentiellen Benutzer eines Spracherkenners recht schwierig, sich zurechzufinden, auch der Wissenschaftler tut sich bereits recht schwer mit der Bewertung der verschiedenen Ergebnisse, zumal die beschriebenen technischen Lösungen und Algorithmen stark unterschiedlich sind.

Im Gegensatz zu anderen technischen Systemen kann die Erkennungsleistung von Spracherkennern bisher nicht mit einem objektiven Maß beschrieben werden, sondern es werden von Menschen gesprochene Testlisten verwendet, die dann über ein Erkennungssystem geschickt werden und deren Erkennbarkeit einen ersten, leider nur sehr groben Anhaltspunkt bietet. Dieser Anhaltspunkt ist deshalb so grob, weil bei der praktischen Benutzung eines Spracherkenners das Verhalten des Menschen mindestens ebenso wichtig ist wie die Fähigkeiten des Erkenners selbst.

2. KRITERIEN FÜR DIE LEISTUNGSMESSUNG BEI SPRACHERKENNUNGS-SYSTEMEN

Es ist deshalb zunächst wichtig, sich klar zu machen, welche Einflußfaktoren bei der Benutzung von Spracherkennern bedeutsam und deshalb für eine realistische Bewertung zu berücksichtigen sind. Ganz grob können wir vier Bereiche unterscheiden:

* Der Sprecher:
  Die Stimmqualität des Sprechers, die charakterisiert ist durch Geschlecht, Alter, Dialekt, Sprechgewohnheiten, Stimmvariabilität, pathologische Besonderheiten, Motivation und Ermüdung.

* Die Übertragungsstrecke zwischen Sprecher und Erkenner:
  Die Eigenschaften der Übertragungsstrecke, die sowohl eine elektrische wie auch die akustische Schallübertragung und die Schallwandlung einschließt und die im wesentlichen zu linearen und nichtlinearen Verzerrungen sowie zu Geräuschüberlagerungen führt.

* Die Dialoggestaltung:
  Die ergonomische und aufgabenorientierte Struktur des Sprachdialogs, im wesentlichen beeinflußt durch syntaktische Abläufe, Wahl eines möglichst natürlichen und leicht unterscheidbaren Vokabulars und einen Dialogablauf der Ermüdung verhindert. Dazu gehört auch die Behandlung von Erkennungsfehlern im Dialogablauf.

* Der Spracherkenner:
  Die technischen Eigenschaften des verwendeten Spracherkenners, beispielsweise gekennzeichnet durch den Erkennungsalgorithmus mit all seinen Einzelheiten, die zulässige Sprechweise (isoliert oder verbunden), die syntaktisch linguistischen Fähigkeiten, die Vokabulargröße, die Erkennungsgeschwindigkeit und letztlich die mit einem möglichst reproduzierbaren Verfahren gemessene Erkennungsrate.

Mit der Messung der Erkennungsrate wollen wir uns nachfolgend näher befassen. Wir sollten jedoch dabei immer bedenken, daß die Erkennungsrate nur ein Teil des sehr viel komplexeren Gesamtproblems der Messung der Leistungsfähigkeit von Spracherkennern ist. Das gilt ganz besonders solange wir keine standardisierten Dialogabläufe beim Einsatz der Spracherkenner verwenden.

Praktisch einsetzbare Erkennungssysteme gibt es derzeit für isoliert gesprochene Einzelwörter und für verbunden gesprochene Wortfolgen/3/. Die Erkennung natürlich gesprochener kontinuierlicher Sprache ist noch nicht anwendungsreif. Die auf dem Markt erhältlichen Erkenner arbeiten nach vergleichbaren und ähnlichen Prinzipien. Zunächst werden aus dem Zeitsignal spektral beschreibbare Parameter gewonnen, die nach empirischen Regeln normiert werden. Manche Spracherkenner besitzen eine explizite Detektion der Wortgrenzen, manche haben die Wortgrenzendetektion in die Worterkennung mit integriert. Hier liegt bereits eine erste Quelle von Klassifizierungsfehlern, wenn beispielsweise bei akustischen Störungen die Wortgrenzen nicht mehr detektiert werden können. Nach der Vorverarbeitung steht ein Wortmuster zur Verfügung, das Ausgangspunkt der eigentlichen Erkennung ist.

Diese Erkennung besteht immer darin, die neuen, unbekannten Signal-

klassen den unterschiedlichen, bekannten Referenzklassen zuzuweisen. Stark vereinfacht können dabei verschiedene Ergebnisse entstehen, die sich aus Abb.1 anschaulich ableiten lassen.

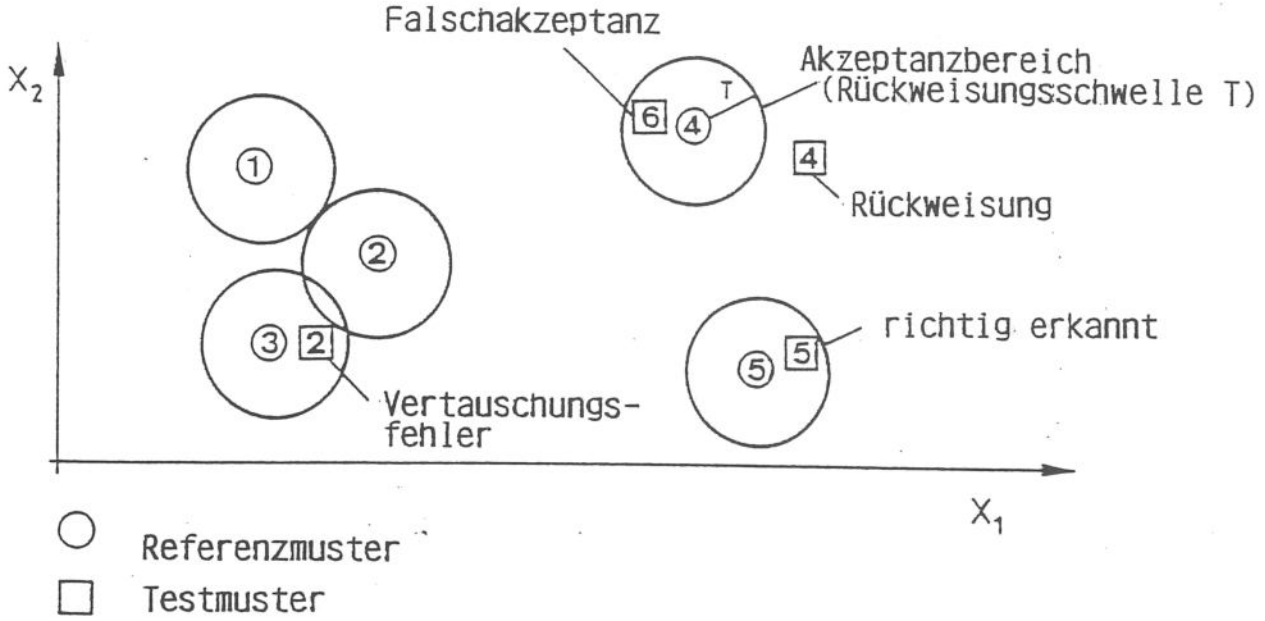

Abb.1: Die möglichen Ergebnisse bei der Musterklassifizierung

In der zweidimensionalen Parameterebene sind die Mittelpunkte der zu erkennenden Klassen 1 bis 5 jeweils von einem Akzeptanzbereich umgeben, der durch die Rückweisungsschwelle T gegeben ist. Dabei können wir unterscheiden:

* Richtige Klassifizierung, d.h. die eingesprochene Wortklasse wird dem richtigen Referenzmuster zugeordnet.

* Falsche Klassifizierung, d.h. die eingesprochene Wortklasse ist zwar gültig, wird aber dem falschen Referenzmuster zugeordnet (Vertauschungsfehler).

* Falschakzeptanzen, d.h. eine im Prinzip zwar nicht gültige Wortklasse (z.B. ein Störsignal) wird einem gültigen Referenzmuster zugeordnet.

* Rückweisungen, d.h. ein gültiges oder ungültiges Eingabemuster wird nicht akzeptiert.

Eine Leistungsmessung von Spracherkennern sollte die Möglichkeit haben, zwischen diesen unterschiedlichen Systemantworten zu unterscheiden. Natürlich setzt ein solches vereinfachtes Modell voraus, daß die einzelnen Testerkennungsvorgänge statistisch voneinander unabhängig sind. Das gilt genaugenommen nur für Isoliertworterkenner ohne syntaktische Verarbeitung. Schon beim Test von Verbundworterkennern spielt die Wortnachbarschaft eine ganz wesentliche Rolle. Oftmals ist es hier auch nicht sinnvoll, sich auf die Einzelworterkennungsrate zu beschränken, da eine brauchbare Eingabe aus mehreren Wörtern besteht und nur die Wortgruppe eine Bedeutung hat (beispielsweise die Bestellnummer in einem Bestellsystem).

Ein systematischer Test sollte natürlich alle Eigenschaften des Spracherkenners unter den möglichen praktischen Einsatzbedingungen

erproben. Dazu gehört beispielsweise die Empfindlichkeit gegen unterschiedliche Sprechlautstärke, Sprechgeschwindigkeit und Sprecher oder auch die Erkennungssicherheit bei erhöhter Geräuschbelastung. Alle diese Parameter lassen sich nur sehr marginal testen, da bisherige Testsysteme solche Besonderheiten nur eingeschränkt abdecken.

Die bisher vorgeschlagenen und praktisch verwendeten Verfahren können wir im wesentlichen in zwei Kategorien einteilen:

* Statistisch objektivierbare Tests, bei denen mit Hilfe von umfangreichen, von Menschen gesprochenen Wortlisten Erkennungsversuche gemacht werden.

* Tests, bei denen objektiv definierbare Abstandsmaße verwendet werden.

3. LEISTUNGSMESSUNG MIT NATÜRLICHEN TESTDATEN

Der naheliegendste Test für Spracherkenner besteht darin, den Erkenner in seiner Anwendungsumgebung und mit seiner Anwendungsaufgabe zu erproben. Da das natürlich nicht für jede mögliche und denkbare Anwendung gemacht werden kann, gibt es inzwischen Tests, die Listen gesprochener Wörter verwenden, die den potentiellen Anwendungen am ehesten entsprechen und mit denen versucht wird, die Erkennungsrate unter möglichst wirklichkeitsnahen Bedingungen zu ermitteln /4,5,6/. Solche Tests sollten standardisiert sein. Nur dann sind Ergebnisse einigermaßen vergleichbar. Erste Ansätze zu einer solchen Standardisierung laufen seit einigen Jahren, unter anderem auch in dem noch zu besprechenden europäischen Projekt SAM /7/.

Es gibt inzwischen eine Fülle von Testdatenbasen, die entweder von Benutzern, von Herstellern oder auch von unabhängigen Organisationen zusammengestellt worden sind /4,6,7/. Manche dieser Datenbasen sind unter gewissen Bedingungen frei verfügbar und werden gegen Erstattung von Kosten für Datenmaterial und Kopieren auch fremden Interessenten zugänglich gemacht.

Allerdings decken die Datenbasen keineswegs alle praktisch vorkommenden Anwendungsfälle ab, sondern repräsentieren häufig eher zufällige Zusammenstellungen von Ausschnitten potentieller Anwendungen. Variationen der verwendeten Sprecher sind dabei ebenso wie die Aufnahmebedingungen mehr oder weniger zufällig.

Eine der ältesten Datenbasen wurde vom US Postal Service bereits Anfang 1980 entwickelt/8/, als man dort erstmals Pakete mit Spracheingabe sortieren wollte. Hier ging es darum, für die sprecherabhängige Spracherkennung unter hoher Geräuschbelastung bei kleinem Wortschatz (10 Ziffern und 10 Steuerwörter) Testmaterial zu sammeln. Man entschied sich deshalb dafür, von 10 Sprechern jeweils 80 mal das ganze Vokabular sprechen zu lassen und erhielt so eine recht große Stichprobe, bei der die Ziffern verbunden als Zifferngruppen gesprochen wurden.

Aus neuerer Zeit stammt die umfangreichste Testliste für sprecherabhängige Erkennung, welche von der japanischen elektrotechnischen

Industrie aufgenommen wurde und bei der von 150 Sprechern jeweils 323 Wörter mehrmals aufgesprochen wurden, so daß die gesamte Stichprobe fast 200 000 Proben umfasst/9/.

Seit einiger Zeit sind auch Bemühungen in Gang gekommen, ähnliches Testmaterial für Erkenner für kontinuierliche Sprache bereitzustellen, das dann bereits für die Entwicklung solcher Systeme Verwendung finden kann. Im Rahmen mehrere Projekte wird in den USA derzeit eine Datenbasis erstellt, bei der von 630 Sprechern mit gezielt ausgewählten Dialekten und Spracheigenheiten jeweils 10 Sätze gesprochen werden. Diese Sätze sind nach phonetischen Gesichtspunkten unter Benutzung der häufigsten Wörter der englischen Sprache ausgewählt. Eine zweite Testliste für kontinuierliche Sprache, ebenfalls in den USA erstellt, enthält 2000 Sätze einer graphischen Datenbasis, gesprochen von 160 Sprechern für den Test der Sprecherunabhängigkeit und von 12 Sprechern zum Test sprecherabhängiger Systeme.

Das wesentliche Problem bei der Anwendung von Worttestlisten besteht in dem enormen Testaufwand, der nötig ist um auch nur einigermaßen vergleichbare und aussagekräftige Ergebnisse zu erhalten. Das gilt auch dann, wenn bereits Einschränkungen hinsichtlich des Anwendungsbereichs und des Benutzerverhaltens möglich sind.

Unter der Voraussetzung, daß über die Wahrscheinlichkeitsverteilung der Erkennungsraten Annahmen zulässig sind, läßt sich beispielsweise die Zahl der Tests angeben, die für eine vorgegebene Sicherheit der Aussage zur Fehlerrate nötig ist. So ist etwa bei einer zu erwartenden Fehlerrate von 0,1 Prozent und einem Vertrauensintervall von 95 %, in dem die zu erwartende Fehlerrate mit 90-prozentiger Sicherheit liegt eine Teststichprobe von 400 000 Mustern erforderlich /10/. Es ist völlig klar, daß die Stichprobe umso grösser sein muß, je kleiner die zu erwartende Fehlerrate eines Erkennungssystems ist. Die heute allgemein verfügbaren Testlisten erfüllen durchweg nicht diese strengen Kriterien. Aber auch wenn es Testlisten solchen Umfangs gäbe, würde die Durchführung eines Tests damit einen kaum zu erbringenden Aufwand erfordern. Es ist also in der Regel nötig, sich mit erheblich kleineren Teststichproben zu begnügen und ergänzende Untersuchungen einzubeziehen.

Solche ergänzende Information bietet beispielsweise die Auswertung von Verwechslungsmatritzen, die recht gute diagnostische Hinweise auf problematische Erkennungsfehler geben können. Die Abb.2 zeigt, daß die Angabe eines mittleren Erkennungsfehlers aus durchaus verschiedenen Erkennungsleistungen herrühren kann und daß deshalb die sorgfältige Auswertung von Verwechslungsmatritzen eine Hilfe sein kann. Dabei muß natürlich bei der Erstellung der Verwechslungsmatrix davon ausgegangen werden, daß die Zuordnung von gesprochener und erkannter Klasse bekannt ist, was zumindest bei kontinuierlicher Sprache nicht immer ganz leicht ist. Wenn beispielsweise die Zifferngruppe 345 gesprochen wurde und es wurde 45 erkannt, so gibt es theoretisch die Möglichkeit, daß lediglich die 3 nicht richtig erkannt wurde, es hätte aber auch die 4 mit der 3 und die 5 mit der 4 verwechselt worden sein können. Dabei wäre dann etwa die 5 überhaupt nicht erkannt worden usw.

a)

		erkannt 1	2	3
	1	0.33	0,33	0,33
gesprochen	2	0,33	0,33	0,33
	3	0,33	0,33	0,33

b)

		erkannt 1	2	3
	1	0,9	0	0,1
	2	0,1	0	0,9
	3	0	0,9	0,1

Abb.2: Verwechslungsmatritzen für drei Klassen mit den Wahrscheinlichkeiten der Zuordnungen.

Die Fehlerwahrscheinlichkeit eines Erkennungssystems wird üblicherweise berechnet zu

$$p(e) = \sum_{i \neq j} p(Yj/Xi) \cdot p(Xi) \qquad \text{für } i \neq j \qquad (1)$$

wo p(Yj/Xi) die bedingte Wahrscheinlichkeit dafür ist, daß zu der Äußerung Xi das Erkennungsergebnis Yj gehört. Diese Wahrscheinlichkeiten können in einer Verwechslungsmatrix zusammengestellt werden. Danach ergibt sich in der Verwechslungsmatrix der Abb.2a eine Fehlerrate von 67 %, die besonders üble Wirkungen hat, da keinerlei Tendenz der Falscherkennungen zu erkennen ist. In der Matrix der Abb.2b ergibt sich zwar die gleiche Fehlerrate, doch ist hier völlig deutlich, daß offensichtlich eine extreme Verwechslung zwischen den Klassen 2 und 3 stattfindet. Solche typischen Fehler lassen sich dann auch leichter beheben. Wichtig ist zusätzlich noch, daß im Worst-Case-Fall von Abb.2a die Fehlerrate linear mit der Zahl der Klassen steigt.

4. OBJEKTIVIERENDE LEISTUNGSMESSUNG

Statistisch brauchbare Fehlerraten oder Verwechslungsmatritzen lassen sich nur mit einem Aufwand zuverlässig zu messen, der in der Praxis kaum zu erbringen ist. Es ist nur natürlich, daß man sich deshalb schon seit einiger Zeit darum bemüht, einfacher zu bestimmende Qualitätsmaße zu definieren, die noch dazu viel weniger von den Zufälligkeiten der Teststichprobe abhängen.

Dabei spielen zwei Aspekte eine wesentliche Rolle: Eine bestimmte Fehlerrate bedeutet bei einem größeren Vokabular eine höhere Erkennungsleistung und andererseits muß die Bewertung des Schwierigkeitsgrades eines bestimmten Vokabulars in die Beurteilung mit einbezogen werden.

Zu dem ersteren Aspekt wurde mit dem Relative Information Loss RIL ein Maß entwickelt, das von der Modellvorstellung ausgeht, daß ein Spracherkenner letztlich wie ein Übertragungssystem behandelt werden kann, das am Eingang Information erhält und am Ausgang irgendwelche veränderte Information ausgibt/11/.

Die Information am Eingang ist dann durch die Entropie gegeben:

$$H(X) = -\sum_{i} p(Xi) \cdot \log p(Xi) \qquad (2)$$

Die Äquivokation des gedächtnisfreien Kanals

$$H(X/Y) = -\sum_{i,j} p(Xi,Yj) \cdot \log p(Xi/Yj) \qquad (3)$$

d.h. die Entropie des Kanaleingangs bei bekannten Ausgangswahrscheinlichkeiten kann als Maß für den Informationsverlust auf dem Kanal interpretiert werden. Der RIL ist dann die auf die Eingangsentropie normierte Äquivokation:

$$RIL = H(X/Y) \;/\; H(X) \qquad 0 \leq RIL \leq 1 \qquad (4)$$

Sie stellt ein Maß dar, das unabhängig von der Vokabulargröße eine Leistungszahl für einen Erkenner liefert, wobei die in der Verwechslungsmatrix gemessene Daten ausreichen. Der RIL liegt für den schlechtesten Fall (Abb.2a) beim Wert 1 und für ungestörte Erkennung bei 0. In Abb.3 ist die Auswertung eines Tests verschiedener Spracherkenner dargestellt, bei denen sowohl die mittlere Fehlerrate als auch der RIL in Abhängigkeit vom Systempreis dargestellt wurde. Die verschiedenen Kurven wurden mit unterschiedlichen Datenbasen gemessen, die aber alle aus kleinem Vokabular (10 Ziffern und wenige Steuerwörter) bestanden. Die Erkenner wurden nur mit einzeln gesprochenen Wörtern getestet, obwohl die teureren Systeme verbunden gesprochene Wortfolgen erkennen konnten. Man sieht, daß einerseits mit unterschiedlichen Tests doch recht unterschiedliche Erkennungsergebnisse erhalten werden, daß aber andererseits der Verlauf der Kurve des RIL recht gut mit der Gesamttendenz harmoniert.

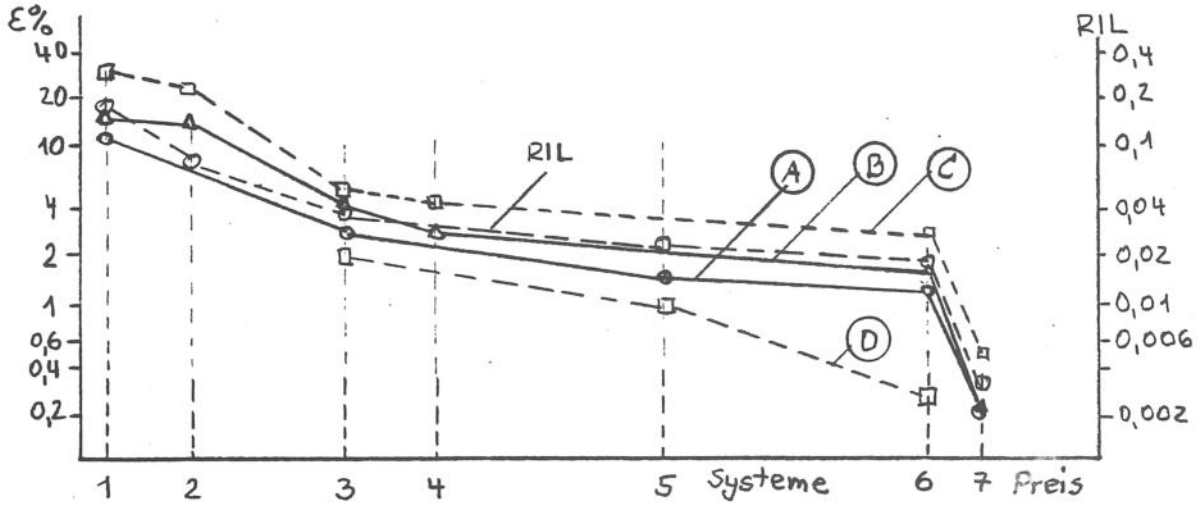

Abb.3: Fehlerrate und RIL in Abhängigkeit vom Systempreis, gemessen mit verschiedenen Testdatenbasen (A,B,C,D)

Ein anderes, sehr wesentliches Problem ist die Beurteilung des Schwierigkeitsgrades der Erkennung eines bestimmten Vokabulars. Es ist sofort einzusehen, daß etwa die Unterscheidung der Wörter "leben" und "heben" erheblich schwieriger ist als die der Wörter "leben" und "Natur". Der phonetische Abstand der Wörter ist im zweiten Fall größer. Besonders wichtig ist es, bereits bei der Zusammenstellung eines bestimmten Vokabulars für eine Erkennungsaufgabe eine Vorstellung von den zu erwartenden Fehlermöglichkeiten, also eine a-priori-Fehlerwahrscheinlichkeit zu definieren.

Besonders wichtig ist eine solche Vorhersage bei großen Vokabularien, bei denen eine angenäherte Abschätzung der möglichen Komplikationen auf Grund empirischer phonetischer Betrachtungen kaum möglich ist. Es wurden deshalb Verfahren entwickelt, die auf der Basis phonetischer Transkriptionen der zu erkennenden Wörter mit Hilfe der dynamischen Programmierung phonetische Abstandsmaße ermitteln, die einen ersten Hinweis auf mögliche Probleme geben.

Auf einem menschlichen Worterkennungsmodell baut das in /12/ beschriebene Verfahren auf. Im wesentlichen werden bei dieser Modellvorstellung zwei Parameter ermittelt, der Signal-Störabstand der nötig ist, um eine vergleichbare menschliche Erkennungsleistung zu erreichen und die Abweichung der Verwechslungsmatritzen von automatischem und menschlichem Erkenner. Der schematische Ablauf ist in Abb.4 dargestellt.

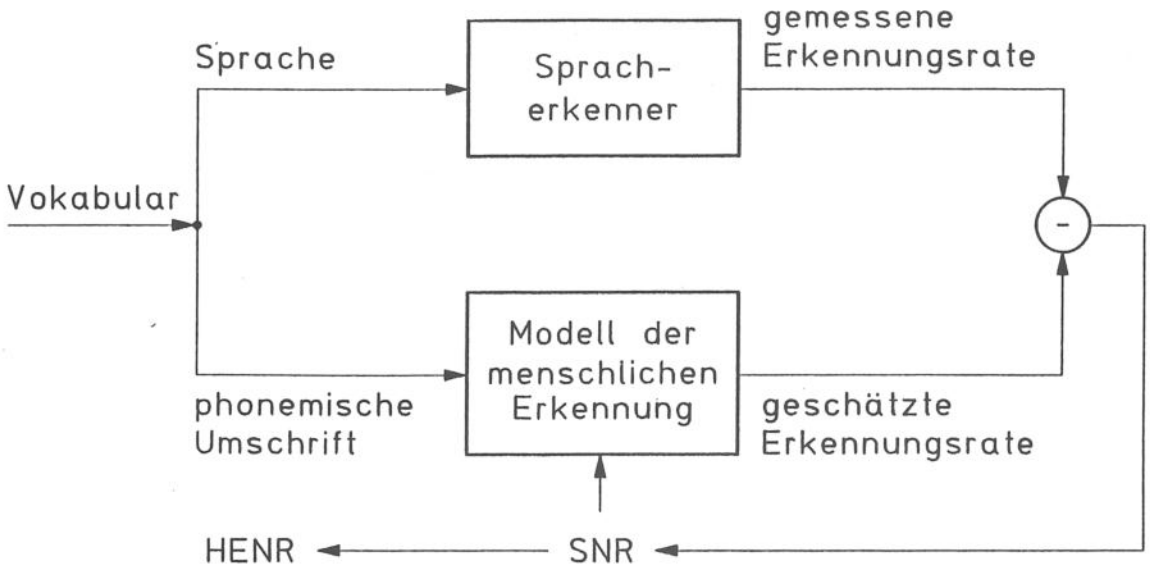

Abb.4: Schematische Darstellung der Bestimmung des Human Equivalent Noise Ratio HENR

Anhand von Vergleichsmessungen wird ein Human Equivalent Noise Ratio HENR gemessen. Hohes HENR bedeutet geringere Leistungsfähigkeit des Erkenners, d.h. der Mensch/Erkenner benötigt zum Erreichen einer bestimmten Erkennungsrate einen hohen Signalstörabstand. Das HENR ist für verschiedene Testvokabularien und den gleichen Erkenner konstant, wenn vorausgesetzt werden kann, daß der Erkenner nach ähnlichen Prinzipien arbeitet wie der menschliche Beobachter. Um diese Annahme weiter zu verfizieren werden die Verwechslungsmatritzen von automatischem und menschlichem Erkenner miteinander verglichen und daraus ein Korrekturfaktor bestimmt. Unterschiedliche Erkenner lassen sich so recht einfach vergleichen. Die umfangreichen Tests mit Versuchspersonen sind nur einmal zu machen.

Einen anderen, aber durchaus vergleichbaren Ansatz verfolgt die Berechnung der Effective Vocbulary Capability EVC/13/. Die EVC gibt die maximal mögliche Vokabulargröße an, die ein Erkenner bei Einhaltung einer bestimmten Fehlerrate noch verarbeiten kann. Abb.5 zeigt das Prinzip. Hier wird in einer Art umgekehrter Definition statt der Fehlerrate das bei Einhaltung einer bestimmten Fehlerrate maximal zulässige Vokabular bestimmt. Letztlich wird ein Wahrschein-

lichkeitsmodell für die Abschätzung des Fehlerverhaltens eines Erkenners entwickelt, das auf einer Bestimmung der a priori-Verwechslungswahrscheinlichkeit beruht.

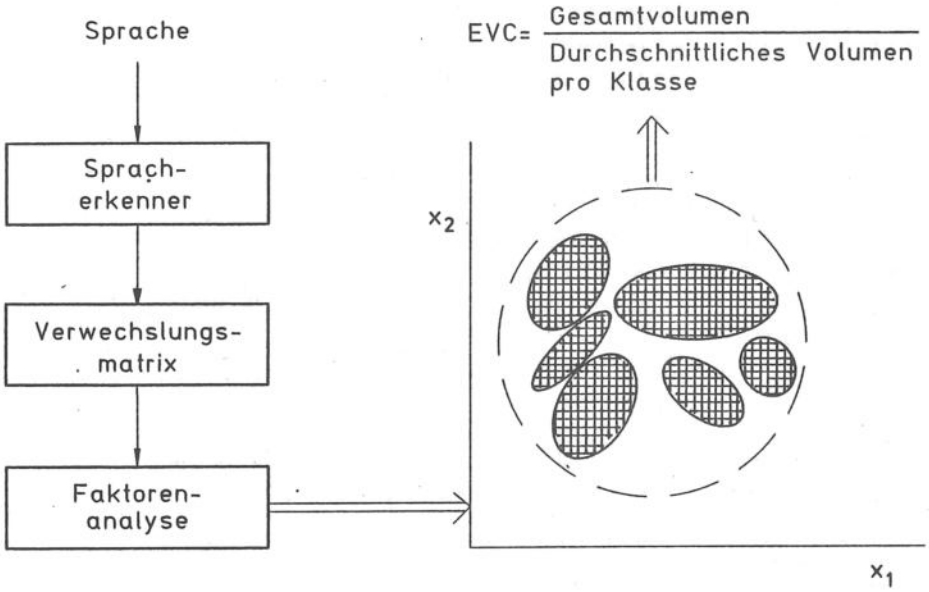

Abb.5: Prinzip der Definition der Effective Vocabulary Capability EVC.

Alle objektivierenden Testverfahren messen nicht allein die Fehlerrate oder auch beispielsweise Verwechslungsmatritzen, sondern sie messen die dem Benutzer normalerweise nicht zugänglichen Abstands- und Wahrscheinlichkeitsmaße, die eine zuverlässigere Abschätzung der Erkennerleistung erlauben weil dabei viel stärker auf die Funktionsweise der Erkenners Rücksicht genommen wird.

Eine noch weitergehende Form der Objektivierung wird schließlich durch Verwendung von synthetischer Sprache als Testsignal erreicht, die in ganz bestimmter Form verändert wird. In /14/ werden dazu künstliche Wörter der Struktur CVC (Konsonant-Vokal-Konsonant) in Form von Minimalpaaren durch Analyse-Synthese-Verfahren in einer Art variiert wie sie auch bei natürlichen Sprechern beobachtet werden kann. Im Gegensatz zur Verwendung natürlicher Testmaterialien lassen sich so gezielt einzelne Eigenschaften des Sprachmaterials verändern und damit die speziellen Eigenschaften des zu untersuchenden Erkenners ausloten. Die Auswertung kann dann wieder mit einer der schon beschriebenen Methoden geschehen.

Die meisten Testsysteme für Spracherkenner haben große Ähnlichkeit mit bereits eingeführten Verfahren zur Messung der Sprachverständlichkeit bei Übertragungssystemen. Das ist nur natürlich, denn ein Erkenner kann letztlich mit einem Menschen verglichen werden, der Sprache erkennen soll, die beispielsweise über das Telefon übertragen wurde. Für solche Verständlichkeitsmessungen wurden in den letzten Jahren recht brauchbare objektive Meßverfahren entwickelt, die Modellvorstellungen der menschlichen Perzeption verwenden. Einstweilen sind solche Verfahren nur auf die Eigenschaften des peripheren menschlichen Gehörs beschränkt. Mit vertiefter Kenntnis der menschlichenn Sprachperzeption werden in Zukunft auch objektive Messungen an Erkennern für kontinuierliche Sprache oder sogar an sprachverstehenden Systemen möglich sein.

5. DAS ESPRIT-PROJEKT SAM

Fest vereinbarte und genormte Testverfahren für Spracherkenner sind derzeit noch nicht vorhanden, wenngleich weltweit intensive Bemühungen im Gange sind, zu verbindlichen Standards zu kommen. In den USA unterhält das National Bureau of Standards seit fast 10 Jahren bereits eine spezielle Arbeitsgruppe und das IEEE hat bereits vor einigen Jahren eine Working Group on Speech I/O Performance Assessment gegründet. In England wurde eine Speech Technology Assessment Group STAG gebildet, an der die Industrie und verschiedene staatliche Stellen beteiligt sind. Schließlich beschäftigt sich die COST als Organisation der Postverwaltungen in der Working Group 209 seit einigen Jahren mit der Leistungsmessung von Spracherkennern. Im Rahmen des Forschungsprogramms ESPRIT der Europäischen Gemeinschaften gibt es nun seit zwei Jahren ebenfalls ein Gemeinschaftsprojekt vieler Firmen und Institutionen zur Leistungsmessung von Sprachsystemen, das Projekt SAM (Multilingual Speech Input/Output Assessment, Methodology and Standardisation)/7/.

Das Projekt hat sich drei Arbeitsbereiche vorgenommen, den Test von Erkennern, von Sprachsynthesesystemen und die Entwicklung und Bereitstellung der entsprechenden Werkzeuge wie Datenbasen, Testverfahren und Arbeitsplatzsystemen. In einem ersten Ansatz wurden Datenbasen für den Test von Worterkennern auf einer optischen Disk EUROM.0 bereitgestellt. Damit steht für fast alle europäischen Sprachen eine von der Struktur her einheitliche Datenbasis zur Verfügung, die im wesentlichen aus einzeln oder verbunden gesprochenen Zifferngruppen besteht. Für die Anwendung dieser Datenbasis und den Test selbst wurde die Workstation SESAM entwickelt, die aus einem PC besteht, der mit entsprechenden Ein/Ausgabekanälen versehen ist, so daß die Daten der Disk in einen zu testenden Erkenner eingespielt werden können und das Ergebnis der Erkennung ausgewertet werden kann. Gleichzeitig werden natürlich die entsprechenden Softwaremodule bereitgestellt.

In ähnlicher Weise werden beispielsweise einheitliche Sprachlisten für die Messung der Verständlichkeit von synthetischer Sprache bereitgestellt. Systematisch konstruierte "sinnlose" Sätze, die nach einheitlichen Regeln erstellt werden, garantieren, daß die Ergebnisse solcher Messungen besser vergleichbar werden.

SAM ist ein wichtiger Schritt zu einer internationalen Standardisierung auf dem Gebiet der Leistungsmessung von Spracherkennern. Zusammen mit den Aktivitäten in den USA und einigen nationalen Aktivitäten in Europa wird hier bereits in naher Zukunft ein wesentlicher Beitrag zum systematischen Vergleich der Spracherkennungstechnologie geleistet.

6. ABSCHLIESSENDE BEMERKUNGEN UND AUSBLICK

Die Messung der Leistungsfähigkeit von Spracherkennern ist nicht nur eine Frage der Festlegung technischer Verfahren, sondern ganz wesentlich auch eine Aufgabe für möglichst internationale Standardisierung. Sprecherabhängige Ganzworterkenner für isoliert oder verbunden gesprochene Kommandos sind im Prinzip Musterklassifikatoren, die sprachenunabhängig arbeiten. Vergleichende Tests mit europäischen Sprachen zeigen, daß zumindest für diese Sprachen keine signifikanten

Unterschiede in der Erkennungsleistung feststellbar sind. Damit lassen sich prinzipiell Tests angeben, die einen internationalen Vergleich von Spracherkennern erlauben. Eine Reihe von bisher zumeist nationalen Standardisierungsgremien beschäftigt sich inzwischen mit der Aufgabe, solche Tests zu definieren. Die Problematik einer rationellen und effektiven Messung der Leistung von Spracherkennern ist damit allerdings noch nicht gelöst. Wichtige Arbeiten sind noch auf folgenden Gebieten zu leisten:

* Es sind Tests zu entwickeln für Isoliertworterkenner und Verbundworterkenner, die nicht nur Fehlerratenmessungen erlauben, sondern auch gezielte diagnostische Hinweise auf Stärken und Schwächen eines Gerätekonzeptes erlauben.

* Ein besonders komplexes und ungelöstes Problem scheint die Beurteilung von sprecherunabhängigen Erkennern zu sein. Hier ist Grundlagenarbeit zu leisten, da bei solchen Erkennern der klassische Test mit Verwendung von Testlisten wegen der gigantischen Größe solcher Testlisten nicht vernünftig zu machen ist - auch wenn bisher sprecherunabhängige Erkenner mit Testlisten aus einigen hundert Sprechern untersucht werden.

* Die für praktische Anwendungen wichtigen Parameter sind systematisch zu untersuchen und ihrer Wichtigkeit nach in Empfehlungen für Systembeurteilungen zusammenzustellen, d.h. Erkennertests müssen auch auf die Anwendungsaspekte bezogen werden.

* In der komplexen Aufgabenstellung der Mensch-Maschine-Kommunikation lassen sich der Erkenner und das Anwendungsumfeld recht gut beschreiben. Eine präzise Definition des Benutzerverhaltens ist jedoch nur schwer möglich. Hier sind weitere Grundlagenuntersuchungen erforderlich.

* Die Möglichkeiten von Verfahren zur objektiven Messung von Leistungskriterien für Spracherkenner sind weiter zu entwickeln. Diese Verfahren werden in Zukunft vermutlich unerläßlich sein zur Beurteilung von Erkennern für kontinuierliche Sprache und schließlich auch für sprachverstehende Systeme.

7. Literatur

/1/ Lea, W.A.: What causes speech recognizers to make mistakes? Proc. of ICASSP 1982, Paris, 2030-2033

/2/ Verschiedene Autoren in Proc. ICASSP 90, Albuquerque, New Mexico April 1990

/3/ Class, F., Katterfeldt, H., Regel, P.: Methoden und Algorithmen der Worterkennung. Informationstechnik it, (31), 5/89, 306-315

/4/ Mangold, H.: Zur Messung der Leistungsfähigkeit von Spracherkennungssystemen. Informatik Fachberichte 125, 1986, Springer-Verlag, 104-117

/5/ Pallett, D.S., Fisher, W.M., Fiscus, J.G.: Tools for the Analysis of Benchmark Speech Recognition Tests. Proc. Int. Conf. Acoust. Speech and Signal Proc. ICASSP 1990, Albuquerque, 97-100

/6/ Mangold, H.: Assessment of Speech Recognizers in Public Information and Ordering Systems. Proc. ESCA Workshop and Tutorial Day on Speech Input/Output Assessment and Speech Databases, Nordwijkerhout, Netherlands 1989

/7/ Fourcin A.J., Harland G., Barry W.,Hazan V.: Speech Input and Output Assessment - Multilingual Methods and Standards. Ellis Horwood, Chichester, England, 1989
/8/ Baker J.M., Pallett D.S., Bridle J.S.: Speech Recognition Performance Assessments and Available Databases. Proc. Int. Conf. Acoust. Speech and Sig. Proc. ICASSP 1983, Boston
/9/ Itahashi S.: A Japanese Language Speech Database. Proc. of ICASSP 1986, Tokyo, pp.7.4.1-4
/10/Baker, J.M.: How to measure up: Testing speech recognisers. Proc. of the Workshop on Standardisation for Speech I/O-Technology. Nat. Bureau of Standards, Gaithersburg, 1982
/11/Woodard, J.P., Lea, W.A.: New Measures of Performance for Speech Recognition Systems. Proc. of ICASSP 1984, Denver, pp.9.6.1-4
/12/Moore, R.K.: Evaluating Speech Recognizers. IEEE Trans. Acoust. Speech and Signal Proc., ASSP-25, no. 2, April 1977, pp. 178-183
/13/Taylor M.M.: Issues in the Evaluation of Speech Recognition Systems. J. Am. Voice I/O Soc. Vol.3, 1986, 34-68
/14/Steeneken H.J.M., van Velden J.G.: Objective and Diagnostic Assessment of (Isolated) Word Recognizers. Proc. Conf. Acoust. Speech and Sig. Proc. ICASSP 1989, 540-543