PROGRAMM UND ERGEBNISSE DER AUTOMATISCHEN SPRACHVERARBEITUNG AN DER TECHNISCHEN UNIVERSITÄT DRESDEN

Walter Tscheschner

Technische Universität Dresden

1. Arbeitsgruppe (AG) Sprachkommunikation

Im Rahmen der Bildung der Sektion Informationstechnik an der Technischen Universität Dresden wurde im Herbst 1969 die AG Sprachkommunikation im Bereich Kommunikation und Meßwerterfassung, zusammen mit den AG Akustik, Meßwerterfassung und Meßtechnik, installiert. Dieser Bereich war aus dem ehemaligen Institut für Akustik hervorgegangen und wurde entsprechend seiner Tradition später wieder als Bereich für Akustik und Meßtechnik bezeichnet.

Von der AG Sprachkommunikation wurden in Lehre und Forschung Aufgaben wahrgenommen, die vornehmlich auf dem Gebiet der automatischen Sprachsynthese und der automatischen Spracherkennung lagen und auch heute noch verfolgt werden.

Zu dem Personalbestand der AG gehören zwei Hochschullehrer, vier Assistenten mit überwiegend befristetem Status, zwei Ingenieure für Forschung und Lehre und derzeit fünf Doktoranden, die ein Forschungsstudium bzw. eine Aspirantur absolvieren.

Pro Jahrgang kommen ab dem 6. Semester etwa 10 Studenten in unsere Vertiefungsrichtung, die jedoch je nach Interessenlage schon ab dem 2. Semester die Möglichkeit haben, kleinere Arbeiten bei uns durchzuführen. Es entstehen vielfach relativ feste Bindungen, die sich zumindest auf das erreichbare Ausbildungsniveau sehr positiv auswirken.
Seit 1975 konnten pro Jahr zwei bis drei Dissertationen abgeschlossen werden, und bis heute liegen etwa 250 Diplomarbeiten vor.

2. Motive

Aus dem vielfältigen Motivationsgeflecht, das für wissenschaftliche Arbeiten charakteristisch ist, sollen zwei Aspekte abgehoben werden, die für unsere Arbeit besonders wichtig waren und sind.

Einmal war es die kaum zu unterschätzende Bedeutung der Entwicklung der menschlichen Rechnerkommunikation, insbesondere aus der Sicht der individuellen Beherrschung moderner Informationsströme. Es ergibt sich in zunehmendem Maße eine Situation, daß in volkswirtschaftlichen und vielen anderen Bereichen, ähnlich wie in der Meteorologie, vernünftige Prognosen nur dann effizient sind, wenn generierende Systeme modellierbar und die verfügbaren Rechner und die Rechnerkommunikation so leistungsfähig sind, daß Ergebnisse rechtzeitig bzw. vor einem Wetterereignis vorliegen. Verstehen und Beherrschen vieler Prozesse hängt heute immer mehr mit Rechnerleistung und menschlicher Rechnerbeherrschung zusammen. Gleichzeitig entwickelt sich der operationale Charakter der Informationsverarbeitung progressiv und dieser korrespondiert eng mit sprachlichen Kommunikationsmöglichkeiten.

Als ein zweiter Aspekt wurde gesehen, daß die Sprachkommunikation mit Computern ein Gebiet mit einer hochgradig interdisziplinären Charakteristik ist. Es ist ein Gebiet, das aus sprach- und sprechwissenschaftlichen Disziplinen, aus der Informationstechnik (Nachrichtentechnik, Akustik, Rechentechnik), der Informatik, wie auch der Physiologie, Psychologie, Ergonomie und anderen gespeist wird. Für die ingenieur-technische Ausbildung ist da-

bei von Belang, daß hier die vielfältigen Verflechtungen von Erkenntnisprozessen deutlich gemacht werden können, ein Fakt, der sich insgesamt sehr förderlich auf die Qualifizierung von Denkmodellen auswirkt.

3. Orientierung und Wege

Beiträge zu der hochinnovativen technischen Sprachkommunikation können im wesentlichen über drei Wege erbracht werden. Das sind:

- die Entwicklung von Modellvorstellungen bzw. die Qualifizierung von Prinziplösungen;
- die Entwicklung der technischen Basis für die Forschung und für Systemvarianten;
- die Entwicklung der Applikation, die sich auf Anwendungsbereiche, effiziente Systemkonfigurationen, die Ergonomie und Bedarfsfragen bezieht.

Die hier anzutreffenden Forschungsschwerpunkte verdichten sich zwar zu unterschiedlichen Zeitpunkten, relevante Fortschritte erfordern jedoch die Beachtung der wechselseitigen Beziehungen der verschiedenen Wege.

Die Ableitung des Modells für die biologische Sprachverarbeitung und eine geeignete Projektion auf mathematische Regelwerke kann als eine zentrale Frage der automatischen Sprachverarbeitung angesehen werden. Die Mathematisierung beschränkt sich dabei nicht auf wenige Grundgesetze, sondern sie ist der Art, daß ein hochstrukturiertes Modellkonzept entstehen muß, in dem verschiedene Theorien in unterschiedlichem Maße beteiligt sind. Es haben numerische, strukturelle, symbolische, syntaktische und semantische Beschreibungsebenen miteinander zu korrespondieren, die jeweils unterschiedliche Abbildungsniveaus repräsentieren. Der Verknüpfung der verschiedenen Ebenen miteinander ist dann im Prinzip nur gemeinsam, daß bestimmte Informationsreduktionsprozesse - bei der Erkennung - bzw. Informationsausweitungsprozesse - bei der Synthese - zu absolvieren sind.

Entwurf, Realisierung und Verifizierung von Modellen stützen sich sowohl auf biologisches, wie auch sprachwissenschaftliches Hintergrundwissen. Sie werden über Hardware- und Softwarelösungen untersetzt und sind über Simulationsexperimente zu verifizieren.

Obwohl auch bei der Prinziplösung international unterschiedliche Wege favorisiert werden, wurde von uns versucht, die Komponenten:

. bionische Modelle und Objektbeschreibungsansätze,
. formale Apparate und algorithmische Simulation,
. methodologische Fragen

im engen Konnex zu sehen und bei den einzelnen Arbeiten die jeweiligen Wechselwirkungen zu beachten. Tafel 1 enthält die Aufstellung einiger Arbeiten der 80er Jahre - vornehmlich Dissertationen -, die diesen Ansatz widerspiegeln.

Eine Reihe von Themen, denen im Rahmen des Gesamtkonzeptes eine gewisse integrative Funktion zukam, befaßten sich mit Systemlösungen zur Synthese und Erkennung - [16] s. a. Tafel 2 -. Über sie sollte ein bestimmter Kenntnisstand technisch soweit verdichtet werden, daß ein industrieller Entwicklungsansatz gegeben war.

Hinsichtlich der Qualifizierung der technischen Basis für die Forschungsvorhaben waren wir der Meinung, daß die vielgestaltigen Verarbeitungsprozesse, die bei der natürlichen Sprachverarbeitung vorauszusetzen sind, nur dann effizient erschlossen werden können, wenn eine hinreichend leistungs-

- Integrative bionische Modelle der natürlichen Sprachverarbeitung [1,2]
- Lautheits-Tonheits-Modell für stationäre Vokale [3]
- Modellierung von Komponenten von Erkennungsprozeduren (dynamische Transformation, Strukturlernen) [4,5]
- * Umsetzung digitaler Analysekonzeptionen [6,7]
- Experimentiersystem für die Spracherkennung; Niveau: Worterkennung, kontinuierliche Phrasen, Syntaxsteuerung [8,9]
- * Soft-/Hardwareorganisation in Rechner- bzw. Mehrrechnersystemen für die Sprachverarbeitung [10,11]
- Expertengeführte Dialogsysteme zur Wissensaufbereitung bei Erkennung und Diagnose akustischer Signale [12,13]
- Synthese außereuropäischer Sprachen (Tonhöhensprache / Arabisch) [14,15]

Tafel 1: Arbeiten der AG Sprachkommunikation 1980-90 (Ausschnitt)
(* mit Orientierung auf Forschungstechnik)

1972	SYNI 1	2-Formant-Synthetisator
1973	PBG	Play-Back-Gerät für die Wiedergabe von Sonagrammen zu Demonstrationszwecken
1974	EK I	Worterkenner für 27 Einzelworte auf Großrechner BESM 6
1975	SYNI 2	3-Formant-Synthetisator mit Lochbandsteuerung (Demonstration als Auskunftsterminal für Studentendateisystem)
1977	EK II	Worterkenner für 56 Einzelworte auf Großrechner BESM 6
1977	ROSY 4000	4-Formant-Synthetisator (Syntheseterminal mit Vollsyntheseprogramm auf PR 4000, später auf KRS 4201, ab 1986 auf K 1630)
1979	EK III	Worterkenner mit Vorverarbeitung auf R 300, Lernen und Erkennen (dynamische Programmierung) auf Großrechner BESM 6
1981	EK III/2	Kleinrechnerversion des Erkenners EKIII auf KRS 4201 und K 1620
1984	EK IV	Erkenner mit dynamischer Programmierung für fließend gesprochene Wortfolgen als Experimentiersystem auf K 1620
1987	TUSY	4-Formant-Reihenfilter-Synthetisator in Single-Board-Ausführung für den Einsatz in Forschung, Lehre und Rehabilitation (einschließlich Programmsystem zur reproduktiven Erzeugung und Editierung der erforderlichen Ansteuerdaten auf Kleinrechner K 1630)
1987	AKP	Akustikprozessor (Ergebnis der Zusammenarbeit mit dem Kombinat Musikinstrumente auf dem Gebiet der Psychoakustik)

Tafel 2: Wichtige Prinziplösungen

fähige Rechentechnik als Modellsubstrat zur Disposition steht. Unter unseren Bedingungen konnte jedoch einer solchen Vorstellung nur sehr schwer entsprochen werden. Es bedurfte besonderer Anstrengungen der AG, um hier einen angemessenen Stand zu erreichen. Graphik, Modularität, Rechnerkopplung, Datenbasis, Schaltkreisanwendungen u. a. waren Fragen, die uns neben der eigentlichen Thematik mit z. T. erheblichen Arbeitsanteilen beschäftigt haben.

Eine ähnliche Situation lag ebenfalls bei der Bearbeitung von Aufgaben vor, die auf eine unmittelbare Anwendung abzielten.
Hier waren, bei sehr unzulänglichen technischen Voraussetzungen, Lösungen - vornehmlich auf dem Gebiet der Synthese - bis zum Musterbau voranzubringen. War es doch eine wichtige Frage, auch bei allen Einschränkungen, die innovative Rolle der Sprachkommunikationstechnik praxisrelevant zu machen. Das große Interesse des Blinden- und Sehschwachenverbandes und anderer Partner hat hierbei sehr stimulierend gewirkt. Der Taschenrechner mit phonetischer Ausgabe einschließlich des Spracheditiersystems SPREDI ist in dem Zusammenhang ein hervorzuhebendes Beispiel.

4. Einige Ergebnisse

Als den eigentlichen Ursprung unserer Arbeiten möchte ich den 1957 in Dresden entstandenen Vocoder [17] bezeichnen, der sich auf Arbeiten von Halsey und Swaffield stützt. Mit diesem System konnte nunmehr das Sprachsignal als eine Erscheinung untersucht werden, deren Konstellationen System ↔ Signal ↔ Perzeption auch im Detail zu beobachten waren.

Die mit den Vocoderkanälen gegebenen Möglichkeiten der spektralen Signalbeeinflussung, die sich sinngemäß mit Zeitfiltern, Zeittransformationen, Amplitudenfiltern u. a. m. auf die Analyse zeitlicher u. a. Aspekte ausdehnen läßt, wie auch die Trennung von Anregung und Bewertung, lassen wichtige Eigenschaften des Sprachsignals und der natürlichen Sprachverarbeitung erkennen.
Dazu zwei Beispiele:

1. Über den Vocoder sind verständliche Sätze auch dann zu verstehen, wenn das Pitchsignal monoton ist oder auch aus einer anderen Quelle stammt. Resümee: Das zeitabhängige Kurzzeitspektrum (τ = 40 ms), die Bewertungsfunktion des Ansatzrohres, liefert alle semantisch wichtigen Informationen - gilt z. B. auch für Tonhöhensprachen.

2. Werden geschlossene Silben einer Frequenzbandbegrenzung (Tiefpaß f_g= 750 Hz) und einer Zeitbandpaßbegrenzung (ΔT = 200 ms, Vokaldetektion) unterworfen, dann lassen sich Ergebnisse nach Bild 1 finden. Resümee: Die durch die Formanten F_2, F_3 und F_4 gelieferte Spektralinformation kann, zumindest für die Vokale, nicht sehr erheblich sein. Eingriffe in die Zeitstrukturierung reduzieren dagegen in starkem Maße die Vokalerkennung [18].

Wie spätere Untersuchungen ergaben, resultiert das eigentliche Problem nicht daraus, wie solche Ergebnisse zu gewinnen sind - sie liegen heute bereits in einer sehr großen Anzahl vor -, es ist vielmehr die Frage, wie solche Einzelaussagen zu interpretieren bzw. wie sie in einen größeren Modellrahmen einzuordnen sind. Es zeigt sich, daß auch die so offensichtlichen Aussagen der genannten Beispiele bei einer Modelleinordnung erheblich relativiert werden müssen.
Im 1. Fall ist das Pitchsignal keine so überflüssige Größe, wie es hier den Anschein hat. Vielmehr spielt es für die Setzung der subjektiven Aufmerksamkeit des Perzipienten und die Beschreibung der Lautverbindungsphase eine

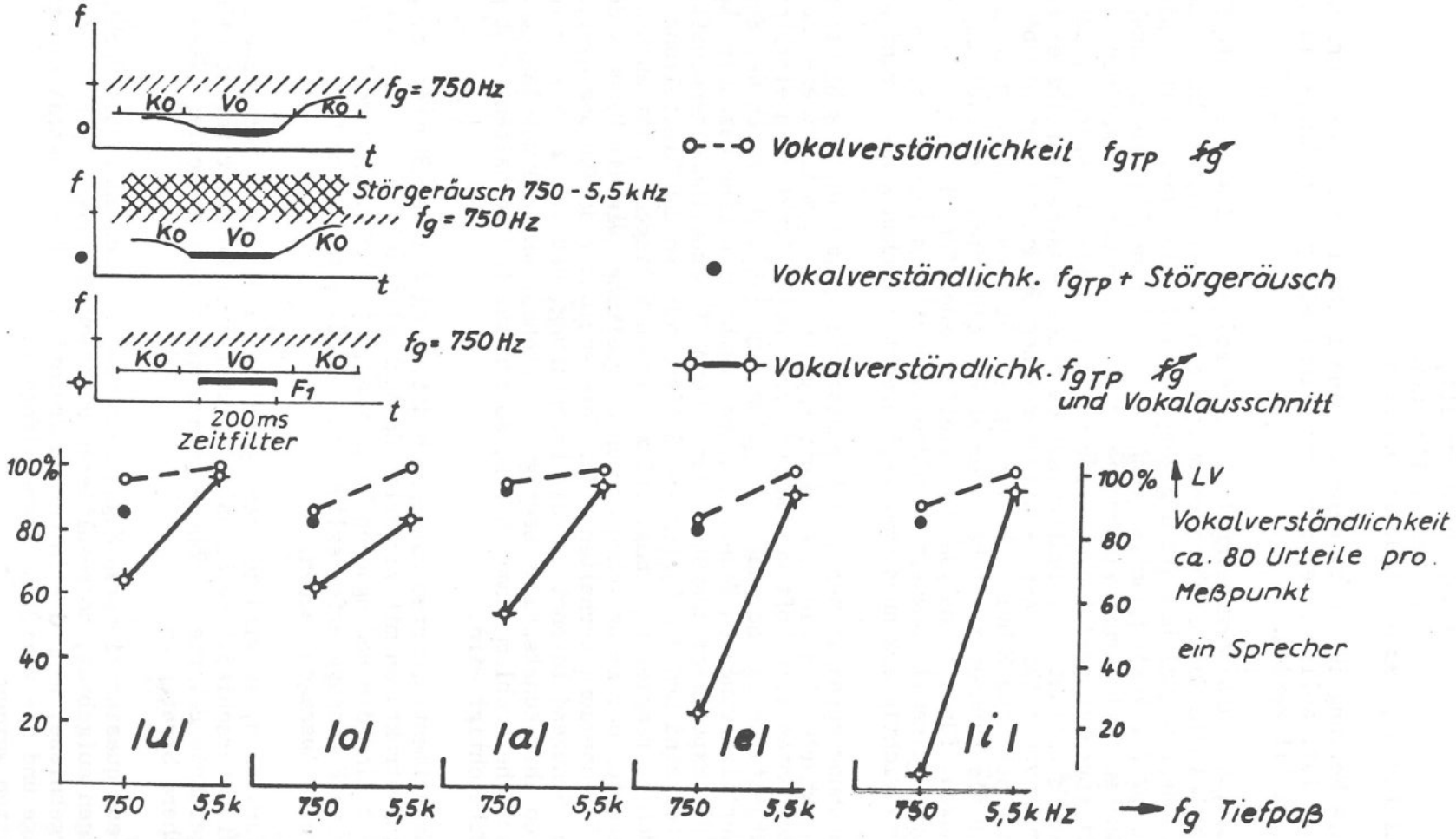


Bild 1 Vokalverständlichkeit in Logatomen bei Begrenzung durch Frequenz- und Zeitfilter

wichtige Rolle. Über dieses Signal kann die Entscheidungsgeschwindigkeit - Zahl der korrekten Entscheidungen pro Zeiteinheit (Akzeptanz) - stark beeinflußt werden. Auch in dem 2. Fall liegt nur ein gewisser Grenzfall vor, der weniger eine reale Sprachsignalsituation beschreibt, als vielmehr die subjektive Fähigkeit, sich an bestimmte Signaleingriffe anpassen bzw. nicht anpassen zu können. Jedoch hier bleibt, daß die Lautverbindungscharakteristik ein dominantes Kriterium ist.
Da m. E. die Wertung der Lautverbindung eine zentrale Frage der Sprachsignalverarbeitung ist, soll sie bei der Betrachtung einiger Ergebnisse in den Vordergrund gestellt werden.

Geläufig ist, daß das Sprachsignal als eine Folge von Zuständen aufgefaßt werden kann, wobei die Zustandsbeschreibungen zumeist mit Spektralabbildungen quasistationärer Signalabschnitte gekoppelt sind. Eine Vorstellung, die sich in dem Konzept der Vektorquantisierung oder in den Realisierungskonzeptionen zur dynamischen Optimierung widerspiegelt. Lösungen zur automatischen Worterkennung mit sprecherspezifischen Referenzmodellen lassen erkennen, daß bei Verwertung der Wortstrukturinformation solche Modellansätze praktikabel sind. Bei der Erkennung von natürlichen, kontinuierlichen Sprachsignalen ergeben sich damit jedoch erhebliche Schwierigkeiten [8, 9]. Sie resultieren im allgemeinen daraus, daß die Parameterzeitcharakteristika keine richtige Bewertung finden. Das gilt sowohl für die Analyse, wie vielfach auch für die Synthese. Maschinen, die die Worte [einemillionzweitausenddreihundert] und [einemilliondreitausendzweihundert] richtig erkennen bzw. zeitgerecht sprechen können, sind selten.

Zur Untersuchung sprachlicher Zeitcharakteristiken und ihrer subjektiven Verarbeitung bewährte sich ein Experimentiersystem, bei dem Erkenner- und Synthetisatorkomponenten verknüpft werden und deren Detailfunktionen einer objektiven und subjektiven Beobachtung zugänglich sind [19]. Bezüglich der Entwicklung und Verifizierung von Modellansätzen ergibt sich eine besondere Leistungsfähigkeit des Experimentiersystems dann, wenn die Funktionseigenschaften genügend variabel sind und das System auf Reizvariationen und Einstellungsveränderungen rasch zu reagieren vermag. Unter solchen Bedingungen ist auch der dynamische Charakter der natürlichen Sprachverarbeitung, wie die Dynamik der Adaption, des Entscheidungsverhaltens, der Referenzmusternachführung usw., gut zu beobachten. Genügend bekannt ist die Erscheinung, daß die Entwickler von Synthetisatoren ihre Kunstsprache sehr gut verstehen, während unbelastete Hörer z. T. erhebliche Probleme damit haben, sofern das Lernverhalten nicht entsprechend berücksichtigt wird.

Wird das Experimentiersystem zur Spracheditierung bei der Synthese eingesetzt, können Signalstrukturen mit variabler Sprachähnlichkeit untersucht werden. Gleichzeitig sind die Bedingungen bestimmbar, die für eine schnelle und sichere subjektive Erkennung erforderlich sind. In diesem Zusammenhang lassen sich zwei generelle Aussagen machen.

1. Fehlt Vorwissen, so wird bei geringer Sprachähnlichkeit individuell vielfach auf unterschiedliche Signaleigenschaften fokussiert (siehe Bild 2 a, das angeführte Beispiel). Höhere Sprachähnlichkeit provoziert dagegen einheitlichere Strategien.

2. Werden aus quasistationären Signalsegmenten mit lautähnlichen Spektren Lautfolgen aufgebaut, so resultieren unterschiedliche subjektive Zuordnungsergebnisse, wenn die Intensitätsrelationen, die spektralen Segmentübergänge und die zeitliche Segmentlänge variiert oder Teile der Folge abgeschnitten werden.

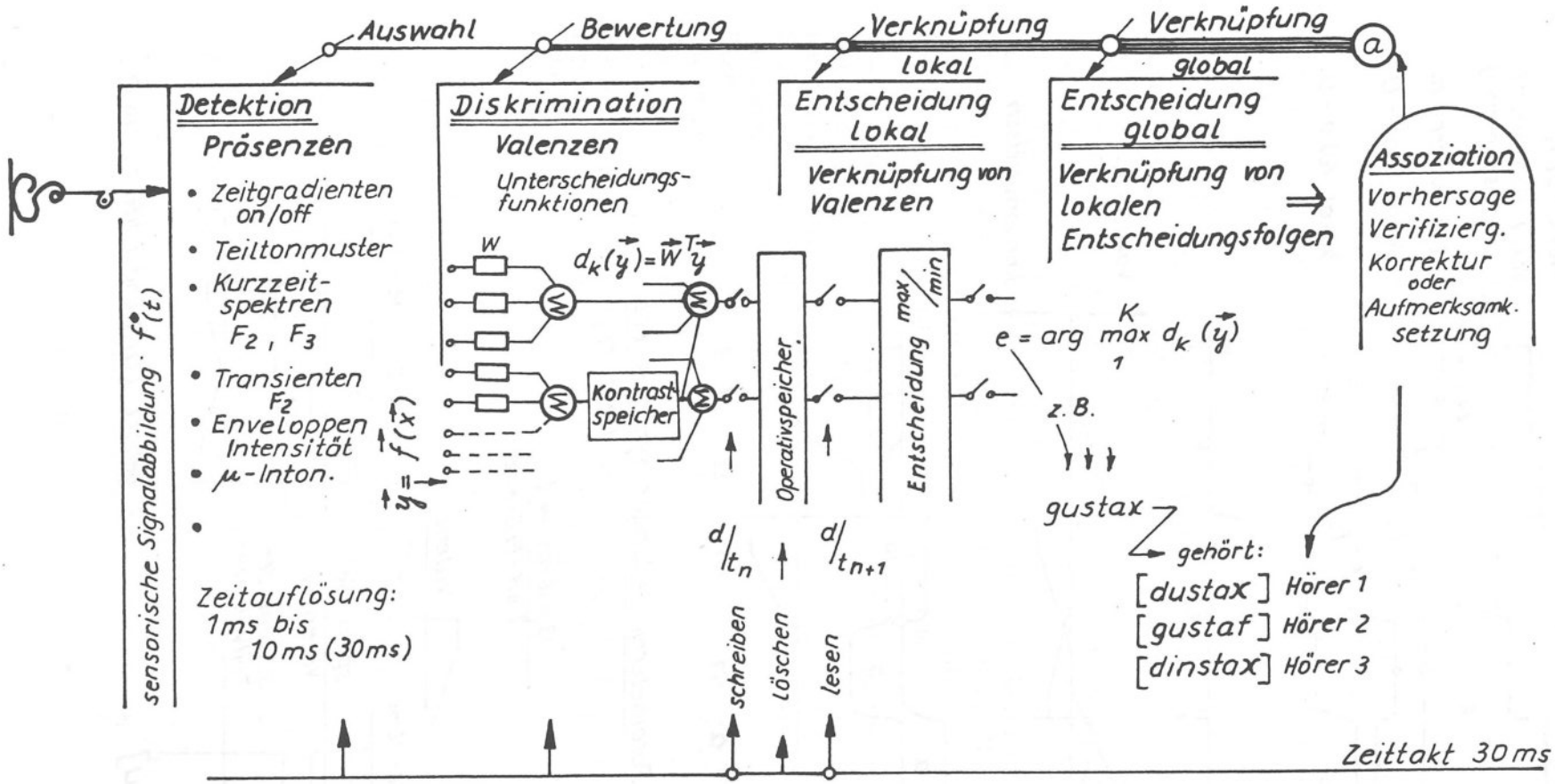
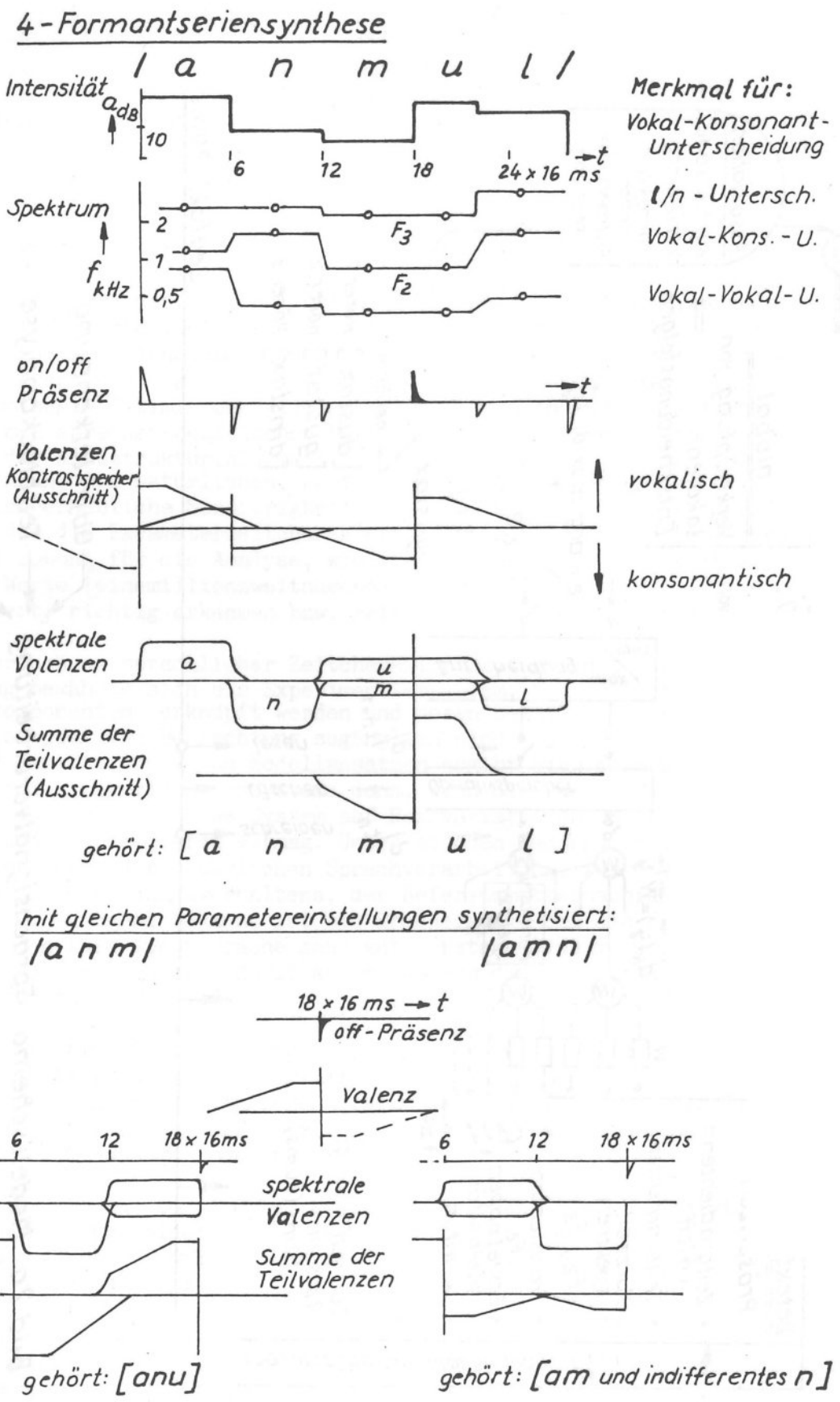


Bild 2a Modellschema Sprachsignalverarbeitung

Objekterkennung
Teilblockanalyse
Spracheditierung

ⓐ Rückführung für Korrekturprozesse individuell sind variable Strategien möglich

Bild 2b **Hörergebnisse und Interpretation** (nach Bild 2a) **bei synthetischer Spektralstruktur /anmul/, /amnul/**



Dazu zwei Beispiele:

synthetisiert: |anmul| gehört [anmul]
Entfernung des Wortteils |ul|, gehört [anu]

|amnul| gehört [amnul]
Entfernung des Wortteils |ul|, gehört [am]

Die Interpretation solcher Ergebnisse der natürlichen Sprachverarbeitung, die sich mit dem Experimentiersystem sicher reproduzieren lassen, war mit ein wichtiges Anliegen der psychophysiologischen Modellansätze in der AG. Zur Erfassung der Phänomene am Lautübergang ist dabei der "Kontrastspeicher" als wichtiges Funktionsprinzip erkannt worden.

Im psychophysiologischen Modellansatz der hierarchischen Sprachsignalverarbeitung [2] ist der Kontrastspeicher [1] der Ebene der Eigenschaftsdiskrimination zuzuordnen - Bild 2a. Er wird aufgerufen durch maximale positive oder negative Gradienten der Intensitätsenvoloppe des Signals, genauer der spektralen Intensitätsenveloppen des Signals (on/off-Präsenzen), und ist durch zwei gegenläufige trapezförmige Unterscheidungsfunktionen beschreibbar. Die Haltezeit ist ca. 2 x 30 ms, und die (frequenzabhängige) Abfallzeit beträgt ca. je 100 ms. Ein on-Sprung betont den Höreindruck Konsonant-Vokal und ein off-Sprung den Höreindruck Vokal-Konsonant. Zusammen mit den spektralen Valenzen - hier Formanteinstellungen hinreichend - sind die resultierenden Höreindrücke der unter Fall 2 genannten Beispiele entsprechend dem Schema in Bild 2b interpretierbar.

Ein weiteres wichtiges Moment der Lautverbindungsphase ist der F_2-Transient. Bei dem verwendeten 4-Formant-Serien-Synthetisator - der quasistationär Konsonanten vom Prinzip her nur unvollkommen modelliert - ist z. B. eine sichere Silbenerkennung für |au| und |am| möglich. Und zwar auch dann, wenn bei sonst gleichem Spektralaufbau der natürliche off-Sprung eliminiert wird. Das einzige Diskriminationskriterium, das erhalten bleiben muß, ist die differierende Zeitdauer für den Formantübergang. Bei $T_{\Delta F_2} > 30$ ms wird für [au] entschieden und $T_{\Delta F_2} < 30$ ms für [am].

Die Kontrastierung durch den Lautwechsel und das Auftreten des vokalischen F_2-Transienten in der Lautübergangsphase ermöglichen es auch, die oben zitierten Vokalerkennungsergebnisse bei Zeitfilteruntersuchungen modellgerecht zu untersetzen.

5. Zusammenfassung

Es werden Arbeitsergebnisse der AG Sprachkommunikation der Sektion Informationstechnik der Technischen Universität Dresden vorgestellt.

Die konzeptionelle Verflechtung spezialisierter Arbeiten zum biologischen und formalen Modellansatz, zur Systembasis (Rechnerverbundsystem, Experimentiersystem) und zur Applikation erlaubt, konstruktive Beiträge zur automatischen Spracherkennung und Sprachsynthese zu leisten.

An einigen punktuellen Ergebnissen wird das erreichte Modellniveau der Sprachsignalverarbeitung angesprochen.

Literatur

1. Hausfeld, H.: Zur zeitlichen Strukturierung des Sprachsignals auf der Grundlage der psychoakustischen Hüllkurvenverarbeitung. Diss. TU Dresden 1984
2. Blutner, F.: Humanes Sprachverarbeitungssystem
   Taschenbuch der Akustik, VEB Verlag TechnikBerlin 1984
3. Ose, R.: Extraktion von Merkmalen zur Vokalklassifikation
   Diss. TU Dresden 1984
4. Weigert, G.: Dynamische Transformation von Merkmalvektorfolgen
   Diss. TU Dresden 1984
5. Kreipe, G.: Ein Experimentiersystem für strukturanalytische Untersuchungen an Sprachsignalen. Diss. TU Dresden 1986
6. Kordon, U.: Eigenschaften digitaler Analyseverfahren für die Sprachverarbeitung. Diss. TU Dresden 1982
7. Kürbis, St.: Ein Experimentiersystem zur Analyse von Sprachsignalen.
   Studientexte zur Sprachkommunikation, H. 1, S. 88 - 96, TUD S 09 1985
8. Paul, V.: Ein Experimentiersystem zur automatischen Erkennung fließender Sprache. Diss. TU Dresden 1985
9. Flach, G.: Syntaktische Steuerung eines Erkennungssystems für fließende Sprache. Diss. TU Dresden 1986
10. Scheide, L.-G.: Heterogenes modulares konzentriert aufgebautes verteiltes Mehrrechnersystem mit verteilter Kontrolle für ein experimentelles sprachakustisches Verarbeitungssystem. Diss. TU Dresden 1986
11. Mittig, A.: Ein Beitrag zur Gestaltung eines modularen Sprachexperimentiersystems. Diss. TU Dresden 1989
12. Haller, K.: Die Anwendung von CAD/CAM zur akustischen Diagnose von Musikinstrumenten. Studientexte zur Sprachkommunikation, H. 5, S. 65 - 70, TUD S 09 1988
13. Langmann, D.: Unscharfe Segmentierung von Sprachsignalen
    Proc. 8. Tagung Akustik, Berlin 1989, S. 131 - 136
14. Pham Hong Ky: Modellierung der vietnamesischen Phonetik mittels Sprachsynthese. Diss. TU Dresden 1984
15. Ali, H.S.: Analytische und technische Untersuchungen zur Synthese arabischer Sprachsignale. Diss. TU Dresden 1989
16. Tscheschner, W.; Hoffmann, R.: 20 Jahre Lehr- und Forschungskollektiv Sprachkommunikation. Studientexte zur Sprachkommunikation, H. 6, S. 5 - 12, TUD S 09 1989
17. Krocker, E.: Aufbau und Untersuchung eines Übertragungssystems für synthetische Sprache. Diss. TH Dresden 1957
18. Tscheschner, W.: Analyse der deutschen Sprache unter besonderer Berücksichtigung der nichtstationären Vorgänge. Diss. TU Dresden 1961
19. Tscheschner, W.: Über ein Experimentiersystem zur automatischen Sprachsignalverarbeitung.
    Bigtech 1988, Berlin, Proc. S. 54 - 59