BMBF

Verbmobil
Verbundvorhaben

Neuere Entwicklungen in der Sprachsynthese

Wolfgang Hess

IKP Universität Bonn

√m **Report 174**
30. September 1996

30. September 1996

Wolfgang Hess

Institut für Kommunikationsforschung und Phonetik
Universität Bonn
Poppelsdorfer Allee 47
53115 Bonn

Tel.: (0228) 7356 - 38
Fax: (0228) 7356 - 39
e-mail: wgh@ikp.uni-bonn.de

Gehört zum Antragsabschnitt: 4

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 D 08 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

Neuere Entwicklungen in der Sprachsynthese[1]

Wolfgang Hess

Institut für Kommunikationsforschung und Phonetik, Universität Bonn
Poppelsdorfer Allee 47, D-53115 Bonn

Als ein Anwendungsgebiet der Sprachsynthese existieren seit vielen Jahren Vorleseautomaten für Blinde (Allen *et al.*, 1987; Kurzweil, 1976; Fellbaum, 1996). Mit einem optischen Lesegerät (Scanner und OCR-Software) zu einem Gesamtsystem integriert, sind solche Systeme heute in der Lage, ihren Benutzern fast jeden beliebigen gedruckten oder maschinengeschriebenen Text vorzulesen. Ein weiterer großer Anwendungsbereich zeichnet sich derzeit ab: automatische Auskunftssysteme über Telefon oder andere Medien, die einen akustischen Ausgabekanal erfordern, z.B. akustische Sprachausgabe im Multimediabereich.

Die meisten Sprachsynthesesysteme verwenden geschriebene Sprache in orthographischer Repräsentation als Eingabe; diese Anwendung wird als *textgesteuerte Sprachsynthese*[2] (*text to speech,* TTS) bezeichnet. Textgesteuerte Sprachsynthese läuft grundsätzlich in drei Schritten ab: 1) Symbolverarbeitung, 2) Verkettung, 3) akustische Synthese. Unter dem Etikett Symbolverarbeitung sind verschiedene Aufgaben zusammengefaßt, beispielsweise Graphem-Phonem-Konversion oder ein Teil der Prosodiesteuerung (Rhythmus und Dauer, Betonung). Am Ausgang der Symbolverarbeitung steht eine Zeichenkette diskreter phonetischer und prosodischer Symbole. Das Verkettungsmodul führt diese in einen kontinuierlichen Strom von Sprachsignalparametern (einschließlich Prosodie) und/oder Artikulationsgesten über; der akustische Synthetisator generiert daraus das Sprachsignal. In manchen Synthesesystemen liefert bereits die Verkettungsstufe prototypische Signale; dann beschränkt sich der Synthetisator darauf, diese zu manipulieren und zu modifizieren.

Im Laufe der Jahre hat sich die Sprachsynthese *bottom-up* von der akustischen Ebene bis hin zu den linguistischen Ebenen entwickelt. Parametrische Analyse-Synthese-Systeme sind seit Dudleys Erfindung des Vocoders vor nunmehr fast 60 Jahren wohlbekannt. Synthesesysteme nach Regeln mit symbolischer (phonetischer) Eingabe, die ein Verkettungs- und ein Synthesemodul benötigen, wurden erstmals in den frühen 50er Jahren entwickelt. In den 70er Jahren konzentrierte sich die Forschung in der Sprachsynthese auf den Bereich der Symbolverarbeitung. Die ersten vollständigen TTS-Systeme kamen gegen Ende der 70er Jahre heraus (Klatt, 1987). Damit war TTS zumindest grundsätzlich möglich. Dies brachte allerdings noch keine generelle Lösung des Problems der Sprachsynthese, sondern bedeutete vielmehr das Ende des ersten Durchgangs einer Iteration. Seit dieser Zeit konzentriert sich die Forschung auf die Verbesserung der Struktur der Sprachsynthesesysteme auf allen Ebenen und auf die Optimierung der Qualität des synthetischen Sprachsignals.

Die primären Anforderungen an eine Sprachsynthese sind *Verständlichkeit* und *Natürlichkeit*. Die Verständlichkeit synthetischer Sprache ist heute der Verständlichkeit natürlicher Sprache schon vergleichbar; die Forderung der Natürlichkeit jedoch ist noch keineswegs erfüllt. Dies dürfte auch der entscheidende Faktor dafür sein, daß die Akzeptanz der Sprachsynthese heute noch zu wünschen übrig läßt. Nach Fellbaum (1996) reagiert ein Benutzer, der normalerweise an Telefonsprache oder an die Sprachausgabe von Rundfunkgeräten gewöhnt ist und unvermittelt mit synthetischer Sprache konfrontiert wird, zumeist ablehnend; dies ändert sich erst dann, wenn der Benutzer weiß, daß sein Kommunikationspartner ein technisches System ist. Die Einsatzmöglichkeiten von Sprachsynthesesystemen sind dementsprechend in der Praxis immer noch vergleichsweise begrenzt, obwohl die Anwendungsgebiete sozusagen vor der Haustür liegen (Sorin, 1994).

Dieser Beitrag soll keinen Überblick über heutige oder historische Sprachsynthesesysteme bieten; dies würde den Rahmen sprengen, und hierfür sei deshalb auf die Literatur verwiesen (z.B. Klatt, 1987; Rühl, 1989; Endres und Wolf, 1980; Endres, 1984; Köster, 1973; Allen, 1992; vgl. auch die Sammelbände Bailly und Benoit, 1992 und Van Santen *et al.*, 1996). Vielmehr sollen Entwicklungen aufgezeigt werden, die zu einer signifikanten Verbesserung der Qualität synthetischer Sprache zumindest im segmentalen Bereich geführt haben. Darüber hinaus werden neue Anwendungsmöglichkeiten im Bereich von Auskunft- und Dialogsystemen diskutiert, die mit Hilfe einer inhaltsgesteuerten Synthese (*concept to speech*, CTS) zu lösen sind, und wo durch Einbeziehen von semantischem Wissen und Domänen-

---

1 Dieser Beitrag war vorgesehen für das EUROFORUM-Symposium "Mensch-Maschine-Schnittstelle in Stuttgart am 11./12.9.1996, das vom Veranstalter dann abgesagt wurde. Eine verkürzte Version dieses Artikels ist unter gleichem Titel als eingeladener Beitrag bei der ITG-Tagung *Sprachkommunikation* in Frankfurt/M. erschienen [in: *Sprachkommunikation*, hrsg. von A. Lacroix. ITG-Fachbericht 139 (VDE-Verlag, Berlin), 89-100]

2 Dieser Terminus und weitere Termini zur Sprachsynthese werden entsprechend dem Entwurf der ITG-Empfehlung (Fachgruppe 4.3.1) zur Terminologie der Sprachakustik (ITG, 1996) verwendet.

wissen eine weitere Verbesserung der Sprachqualität zu erwarten ist.

Der Schwerpunkt dieses Beitrages liegt auf den akustisch-phonetischen Aspekten; die Ebenen der Symbolverarbeitung treten demgegenüber in den Hintergrund. Abschnitt 1 diskutiert die Frage der segmentalen Ebene und beschäftigt sich mit der Wahl der segmentalen Einheiten sowie mit Problemen der Verkettung und der Koartikulation. Abschnitt 2 behandelt die Frage der akustischen Synthese. In beiden Abschnitten stehen regelgesteuerte Verfahren solchen gegenüber, bei denen Elemente natürlicher Sprachsignale unmittelbar zur Synthese herangezogen werden. In Abschnitt 3 wird die Frage der Prosodie behandelt, die bei der Beurteilung der Natürlichkeit synthetischer Sprache eine entscheidende Rolle spielt. Auch hier werden modell- und datengesteuerte Verfahren einander gegenübergestellt. Abschnitt 4 beschäftigt sich mit der Evaluation von Sprachsynthesesystemen, um dann einige Studien zu Synthesesystemen für das Deutsche und ihre Ergebnisse kurz vorzustellen. Abschnitt 5 schließlich ist Aspekten der inhaltsgesteuerten Sprachsynthese gewidmet.

1. Synthese auf segmentaler Ebene – Verkettung und Koartikulation

Das Verkettungsmodul ist das Bindeglied zwischen der niedrigsten symbolischen Ebene und der akustischen Ebene. Eine Kette diskreter phonetischer Symbole, ergänzt durch prosodische Steuerzeichen, wird in einen kontinuierlichen Datenstrom von Sprachsignalparametern oder -abtastwerten transformiert. In der Praxis wird die Verkettung durch einen Satz von Regeln sowie ein Korpus von Sprachdaten gesteuert. Das Korpus kann hierbei in tabellarischer Form empirisches Wissen repräsentieren, beispielsweise Tabellen von Formantfrequenzen; es kann aber auch aus natürlichsprachlichen Daten bestehen.

Die Forschung in diesem Bereich ist gekennzeichnet durch die Interaktion von akustisch-phonetischer Modellierung und Sprachsignalverarbeitung. Phonetik und Phonologie liefert Modelle der Spracherzeugung, regelhafte Beschreibungen für Artikulationsgesten und Koartikulationseffekte sowie die Grundlage für die Wahl der segmentalen Einheiten, während die Signalverarbeitung Verfahren zur Sprachanalyse und -manipulation beiträgt. Kein Modell oder Regelsystem jedoch, so vorzüglich es auch arbeiten mag, ist in der Lage, den Spracherzeugungsprozeß, wie er bei einem menschlichen Sprecher stattfindet, völlig adäquat nachzubilden. Jedes Modul trägt damit systematisch zur Verminderung der Qualität des Ausgabesignals im Vergleich zu einer natürlichen menschlichen Stimme bei. Diese Negativbeiträge so gering wie möglich zu machen, ist eine der Hauptaufgaben.

1.1 Verkettung durch Regeln versus Verkettung mit Hilfe natürlichsprachlicher Daten

Das Verkettungsmodul stellt stets einen Kompromiß zwischen der Zahl und Komplexität der Verkettungsregeln einerseits sowie der Größe und Ausdehnung der Datenbasis andererseits dar. So können wir uns z. B. eine rein regelgesteuerte Architektur denken, die fast ohne natürliche Daten mit einigen Tabellen gespeicherten akustisch-phonetischen Wissens über Sprachsignale auskommt (Anregungsart, Werte für Formanten und Antiformanten, artikulatorische Zielstellungen usw.). Ein solches System berechnet Sprachsignalparameter mit Hilfe eines großen Satzes von Regeln, die die Dynamik der Artikulationsgesten bei der Spracherzeugung nachbilden; je nach Synthetisator werden diese Regeln entweder direkt artikulatorisch formuliert, oder sie sind auf der akustischen Ebene als Funktionen von Parametern in Abhängigkeit von der Zeit realisiert. Im Gegensatz hierzu stützen sich datengesteuerte Systeme auf Bauelemente natürlichsprachlicher Daten (gespeichert in einer parametrischen Darstellung oder direkt als Signale) und kommen mit einem Mindestmaß an Regeln aus. Beide Ansätze besitzen eine Reihe von Vor- und Nachteilen.

- Wie bereits gesagt – kein Modell und kein Regelsystem kann die Qualität natürlicher Sprache erreichen. Dies ist mit das stärkste Argument zugunsten eines datengesteuerten Systems. Ein Synthesesystem, das auf natürlicher Sprache basiert, ist in der Lage, einen großen Teil der natürlichen Sprachqualität zu erhalten.
- Regelgesteuerte Systeme bieten ein Maximum an Flexibilität. Durch systematische Variation der Regeln und einiger weniger Daten ist es möglich, eine große Anzahl von Stimmen oder Sprechsituationen zu realisieren. Die natürlichen Sprachdaten eines datengesteuerten Systems behalten demgegenüber immer einige Spezifika bezüglich des Sprechers bei. Aus diesem Grunde wird es dort schwierig sein, den Sprecher oder die gegebene Stimmqualität zu variieren, es sei denn, das System verfügt über Daten von mehreren Sprechern.

Die Speicherplatzfrage, in früheren Zeiten ein erhebliches Problem, kann heute vernachlässigt werden. Der begrenzende Faktor ist heute der Arbeitsaufwand für die Systemerstellung. In einem regelgesteuerten System betrifft dies die Regeln, die entwickelt, getestet und verifiziert werden müssen. Bei Verwendung natürlichsprachlicher Daten erfordert neben der Datenerhebung, also dem Aufsprechen der Bauelemente, deren Aufbereitung den größten Arbeitsaufwand.

Mit einem regelgesteuerten System können wir aus der symbolischen Eingabe den Sprachparametersatz einschließlich prosodischer Parameter in einem Schritt erzeugen. In datengesteuerten Systemen muß das synthetische Signal erzeugt werden, indem gespeicherte

Segmente verkettet werden, die der Zeichenkette am Eingang entsprechen. Dies erfordert zwei Schritte.

1) Es liegt in der Philosophie datengesteuerter Synthese, daß die einzelnen akustischen Bausteine möglichst wenig modifiziert werden. Es ist beispielsweise nicht beabsichtigt, innerhalb der Segmente über das unbedingt Notwendige hinaus spektrale Modifikationen durchzuführen. Um die prosodische Information zu überlagern, müssen wir allerdings in der Lage sein, die Grundfrequenz, die Dauer und die Intensität des Signals an jeder Stelle zu modifizieren.

2) Um zwei Segmente zu verketten, genügt es üblicherweise nicht, sie einfach aneinanderzufügen. Wir müssen deswegen dafür sorgen, daß in unmittelbarer Nähe der Verkettungsstelle eine spektrale Manipulation möglich ist. Die Verkettungsoperationen können hierbei eine zeitliche wie auch eine spektrale Glättung oder Interpolation erfordern. Kontextabhängige Verkettungsregeln steuern diese Operationen; das größte Problem hierbei ist die Koartikulation.

1.2 Einige Aspekte der Koartikulation

Die Minimierung des Artikulationsaufwandes ist ein wesentliches Prinzip in der zeitlichen Organisation von Artikulationsgesten. Mit einem Minimum artikulatorischen Aufwands soll ein Maximum von Information in das Sprachsignal hineingepackt werden. Die wesentliche kommunikative Randbedingung hierbei ist, daß die Nachricht für den Hörer verständlich sein (und bleiben) soll.

Was ist artikulatorischer Aufwand? Wir haben hierfür kein direktes quantitatives Maß. Plausibel ist jedoch die in der Literatur vertretene These, daß die Maximalgeschwindigkeit einzelner Artikulatoren (Lippen, Unterkiefer, Zunge, Velum) die maßgebliche Größe darstellt (Nelson *et al.*, 1984; hier nach O'Shaughnessy, 1987:112).

Die traditionelle artikulatorische Phonetik sieht jedes Phonem charakterisiert durch eine besondere Einstellung der Artikulatoren, d. h., eine spezifische Zielposition. In der Praxis gilt dies allenfalls für wohlartikulierte Sprache und solche Phoneme, die durch ein (quasi-)stationäres Segment realisiert werden (also Vokale, Nasale, Liquide und Frikative, nicht jedoch Plosive oder Gleitlaute). In kontinuierlicher Sprache werden diese Zielpositionen durch Transitionen verknüpft, die dem Prinzip des minimalen Artikulationsaufwandes unterliegen. Dies heißt zunächst, daß bei gegebener Sprechgeschwindigkeit diese Transitionen so langsam wie möglich ablaufen; aus diesem Grunde benötigen sie einen beträchtlichen Anteil der gesamten Äußerung. Es ist wohlbekannt, daß die Übergänge für das Verständnis der Sprache mindestens ebenso wichtig sind wie die Zielpositionen selbst (Endres, 1973) und in der Sprachsynthese sehr sorgfältig modelliert werden müssen. Diese Tatsache als solche ist heutzutage als beinahe trivial anzusehen; nichtsdestotrotz macht es einen Unterschied in der Sprachqualität aus, ob diese Übergänge durch ein Regelwerk modelliert oder als natürliche Sprachsignalbausteine realisiert werden.

Dies ist nur der einfachste Fall dynamischer Vorgänge in der Artikulation. Koartikulation im eigentlichen Sinn geht weit darüber hinaus. Sprachsignale sind hochredundant; in Erfüllung des Prinzips minimalen Artikulationsaufwandes können daher viele Artikulationsgesten über größere Zeitintervalle hinweg geplant und organisiert werden. Damit werden die artikulatorischen Zielpositionen der einzelnen Phone selbst kontextabhängig. Die Koartikulation ist großenteils vorgeplant (Whalen, 1990); insbesondere werden Artikulationsgesten und Zielstellungen, die für die Realisierung bestimmter Phoneme unerläßlich sind, soweit wie möglich antizipiert; hierdurch können viele Übergänge weiter verlangsamt werden.

Die Prominenz von Koartikulationseffekten in einer Äußerung ist eine Funktion der Lautfolge und damit Funktion der Zeit. Manche Laute erweisen sich als "Koartikulationsschranken"; d. h., der wechselseitige koartikulatorische Einfluß von Lauten über eine solche Schranke hinweg ist relativ gering. Dies gilt immer dann, wenn eine artikulatorische Zielposition voll erreicht und für eine bestimmte Zeit ausgehalten wird. Diese Bedingung erfüllen zunächst Vokale, besonders dann, wenn sie lang oder betont sind. Auch über Nasale oder Frikative hinweg sind koartikulatorische Auswirkungen gering. Umgekehrt sind besonders Verschlußlaute und Gleitlaute anfällig gegen Koartikulationseffekte (Öhman, 1966). Einige Laute, z.B. [r], nehmen eine artikulatorische Extremstellung ein, die mit einem verhältnismäßig hohen Aufwand verbunden ist; hieraus ergibt sich ein beträchtlicher koartikulatorischer Einfluß dieser Laute auf ihre Umgebung.

Daß die Behandlung von Koartikulationseffekten beim Entwurf eines Sprachsynthesesystems eine maßgebliche Rolle spielt und spielen muß, ist damit offensichtlich. In jeder fließenden Rede erwartet der Hörer vom Sprecher diese Koartikulationseffekte; sie zu ignorieren bedeutet eine entscheidende Verschlechterung der Qualität der synthetischen Sprache.

Die Auswirkung des Prinzips des minimalen Artikulationsaufwands auf fließende Rede endet nicht mit der Koartikulation; erwähnt sei nur das Auftreten unbetonter und reduzierter Silben und Wörter, vor allem Funktionswörter in fließender Rede. Reduzierte und deakzentuierte Vokale und Silben werden mit verminderter Dauer und geringerer Gespanntheit der Artikulatoren geäußert und führen deshalb zu zentralisierter Artikulation. Weiterhin werden beim schnellen Sprechen infolge Zeitmangels Zielstellungen von Artikulatoren insbesondere in reduzierten Vokalen nicht mehr erreicht (Lindblom, 1963). Dies führt bis hin zum völligen Verschwinden (Elision) einzelner Laute, im Deutschen insbesondere des Schwa [ə], sowie zu Assimilationen auch über Silben- und Wortgrenzen hinweg. Für das Deutsche sei hierzu insbesondere auf die Arbeiten von Koh-

ler (1990) verwiesen. Sofern diese Phänomene die Folge der realisierten Sprachlaute kategorial verändern, müssen sie im Rahmen der Symbolverarbeitung berücksichtigt werden, während sie im übrigen ihren Niederschlag in den Regeln zur Verkettung bzw. in der Zusammensetzung des Bausteineinventars finden.

1.3 Phonetische Einheiten und Elemente. Das Bausteininventar in datengesteuerten Systemen

Sprachsynthesesysteme benutzen vorwiegend Phone und Allophone, Diphone, Halbsilben oder eine Kombination dieser Einheiten als Basiselemente.

Phone und *Allophone* sind die klassischen Einheiten regelgesteuerter Systeme. Neben dem historischen Aspekt – Sprachsynthese nach Regeln entstand aus stilisierten Spektrogrammen und dem Wissen über Formantfrequenzen von Vokalen und Formantübergängen (Klatt, 1987) – zwang die Speicherplatzfrage die älteren Systeme dazu, Phone zu verwenden. Später wurden Phone dann eingesetzt in Systemen, die bereits über einen wohlentwickelten Regelsatz verfügten, in Anwendungsgebieten, die hohe Flexibilität bezüglich der Zahl der Stimmen oder der Vielfalt von Sprechstilen erfordern (Carlson *et al.*, 1991), oder in multilingualen Systemen, wo zahlreiche Sprachen von der gleichen synthetischen Stimme abgedeckt werden müssen.

Die wichtigsten Eigenschaften von Phonen und Allophonen als Basiseinheiten der Sprachsynthese, ausgedrückt in Vorteilen (+) und Nachteilen (−), sind die folgenden.

+ Phone und Allophone bieten maximale Flexibilität und einen hohen Freiheitsgrad bei der Aufstellung des Regelwerks.
+ Das Inventar der Basiseinheiten ist bei weitem das kleinste (40-50 Elemente).
+ Die Datenrate, die jedes Element benötigt, ist bei weitem die kleinste. Eine Stützstelle oder einige wenige Stützstellen per Baustein genügen, die artikulatorische Zielstellung zu repräsentieren.
− Phone, die nicht mit einer stationären Zielstellung des Vokaltrakts verbunden sind (insbesondere Plosive und Gleitlaute), sind schwierig zu synthetisieren.
− Da die Datenbasis nur Information über stationäre Segmente enthält, muß die gesamte Dynamik der Artikulation einschließlich der Koartikulation in Form von Regeln formuliert und implementiert werden Der Zeitaufwand hierfür ist erheblich (Allen *et al.*, 1987).

Ein *Diphon* bzw. eine *Dyade* ist definiert als der Zeitabschnitt von der Mitte eines Lautes bis zur Mitte des folgenden Lautes (Peterson *et al.*, 1958). Die Verkettungspunkte befinden sich in den stationären Segmenten des Signals, soweit vorhanden. Prinzipiell ist die Zahl der Diphone das Quadrat der Zahl der Phoneme zuzüglich derjenigen Allophone, die für die Synthese relevant sind. Phonotaktische Einschränkungen greifen hier nur wenig; d.h., ein hoher Prozentsatz, im Deutschen z.B. mehr als 75 % (Kohler, 1977) aller möglichen Kombinationen zweier Phoneme kommen in der Sprache tatsächlich vor. Im mehrsprachigen Synthesesystem des CNET (Hamon *et al.*, 1989; CNET, 1991) besitzt die Implementierung für das Französische ungefähr 1600, die für das Deutsche ungefähr 1950 Diphonelemente.

Die wichtigsten Vorteile (+) und Nachteile (−) von Diphonen als Basiseinheiten der Sprachsynthese sind die folgenden.

+ Transitionen sind fast vollständig, Koartikulationseffekte zu einem großen Teil in den Daten enthalten; sie müssen nur noch ergänzend als Regeln formuliert werden.
+ Im Vergleich zu silbischen Einheiten sind Diphone verhältnismäßig kurz, so daß trotz der hohen Zahl von Elementen die Datenbasis vergleichsweise wenig Speicherplatz benötigt.
− Im Vergleich zu Phonen ist die Zahl der Verkettungsstellen im Signal nicht geringer, sondern eher noch etwas größer.
− Insbesondere in Konsonantenfolgen sind manche Koartikulationseffekte nicht adäquat repräsentiert.

Der Einsatz silbischer Bausteine wurde möglich mit der Einführung der *Halbsilbe* (Peterson und Sievertsen, 1960; Fujimura, 1976). Jede Silbe wird in eine initiale und eine finale Halbsilbe aufgespaltet; hierbei entsteht ein zusätzlicher Verkettungspunkt im Silbenkern. Die Anfangshalbsilbe enthält eine initiale Konsonantenfolge, die leer sein kann, und den Beginn des Silbenkerns; die Endhalbsilbe enthält den Rest des Silbenkerns und eine Endkonsonantenfolge, die wiederum leer sein kann.

Die Gesetze der Phonotaktik beschränken die Zahl der Anfangs- und Endkonsonantenfolgen auf einen geringen Bruchteil des durch freie Kombination von Konsonanten prinzipiell Möglichen. Im Deutschen können Anfangskonsonantenfolgen beispielsweise bis zu 3 Konsonanten enthalten, aber nur ungefähr 50 Anfangskonsonantenfolgen sind tatsächlich realisiert (Ruske und Schotola, 1978). Entsprechendes gilt für die Endkonsonantenfolgen, die im Deutschen zwischen 0 und 5 Konsonanten enthalten können, von denen jedoch nur etwa 160 existieren. Für das Englische vorliegende Daten sind dem vergleichbar (Fujimura und Lovins, 1978). In anderen Sprachen mit kürzeren Konsonantenfolgen zwischen benachbarten Silbenkernen werden Halbsilben- und Diphonansatz einander sehr ähnlich.

Die wesentlichen Vorteile (+) und Nachteile (−) von Halbsilben als Basiseinheiten der Sprachsynthese sind die folgenden.

+ Halbsilben sind verhältnismäßig große Einheiten; im Vergleich zu Phonen oder Diphonen wird damit die Zahl der Verkettungspunkte im Signal drastisch reduziert.
+ Halbsilben erfassen die meisten Lautübergänge und eine große Anzahl von Koartikulationseffekten in den Daten. Anfangs- und Endkonsonantenfolgen

sind getrennt; somit sind auch zahlreiche allophonische Variationen in den Bausteinen enthalten.
- Im Vergleich zu Phonemen und Dyaden ist die Zahl der Silben – und damit auch der Halbsilben – in einer Sprache grundsätzlich offen. Manche Fremdwörter oder Eigennamen lassen sich mit einem reinen Halbsilbensystem nicht ohne weiteres synthetisieren (Spiegel *et al.*, 1989).
- Auch Halbsilben decken nicht alle Aspekte der Koartikulation ab; insbesondere nicht in Verbindung mit intervokalischen Plosiven.
- Halbsilben benötigen eine erhebliche Menge an Speicherplatz.

Auch wenn die Zahl der Halbsilben im Vergleich zur Zahl der Silben einer Sprache bereits drastisch verringert ist, enthält ein reines Halbsilbensystem trotzdem noch eine allzu hohe Zahl von Bausteinen (für das Deutsche etwa 5500; vgl. Dettweiler, 1984). In allen realisierten Halbsilbensystemen sind daher Maßnahmen implementiert, die darauf hinauslaufen, durch Abspaltung von Affixen, insbesondere von Suffixen, die Zahl der Kombinationen von Silbenkern und Konsonantenfolge zu reduzieren (Fujimura und Lovins, 1978; Browman, 1980; Dettweiler, 1984; Dettweiler und Hess, 1985). In einigen Systemen führte dies zum durchgängigen Einsatz hybrider Einheiten.

Mit den im wesentlichen phonologisch orientierten Einheiten Diphon und Halbsilbe (letztere mit den zugehörigen Modifikationen) konnte die Qualität der synthetischen Sprache in datengesteuerten Systemen schon entscheidend verbessert werden. Wie jedoch Olive (1990) nachwies, läßt sich eine weitere Verbesserung der Sprachqualität erreichen, wenn das Bausteininventar nach akustisch-phonetischen und weniger nach phonologischen Gesichtspunkten ausgewählt wird. Hierbei ging Olive von einem Diphonsystem aus, das bei kritischen Lautkombinationen um Triphonelemente erweitert wurde. Nach ähnlichen Gesichtspunkten wird auch im CNET-System verfahren (Sorin, 1994).

Das in der neuesten Version des Bonner Systems HADIFIX (*Ha*lbsilbe-*Di*phon-Su*ffix*) ebenso wie in der VERBMOBIL-Synthese SprechMobil eingesetzte Bausteininventar (Portele, 1994, 1996a; Portele *et al.*, 1994) ist ein hybrides (gemischtes) Inventar, nach akustisch-phonetischen Gesichtspunkten zusammengestellt.

Als Grundlage dient ein erweitertes Halbsilbenkonzept. Wie Portele (1994) experimentell nachwies, ist die frühere These, daß Silbengrenzen wie Koartikulationsschranken wirken, nicht haltbar. Koartikulationseffekte sind an Silbengrenzen nahezu ebenso stark wie in der Silbe selbst. Zu diesen Effekten gehören z. B. nasale und laterale Verschlußlösungen, kontextabhängige Entstimmungen von Plosiven und Frikativen sowie weitere Assimilationen von Artikulationsart und Artikulationsort. Daher müssen generell die Silbengrenzen nach akustisch-phonetischen Gesichtspunkten redefiniert werden; in Einzelfällen, wenn dies nicht ausreicht, sind eigene Bausteine zu definieren.

In diesem Inventar sind die folgenden Elementetypen (jeweilige Anzahl in Klammern) enthalten (Portele, 1996a).

- Anfangshalbsilben (1086): "klassische" Anfangshalbsilben, zusätzlich: Kombinationen mit Plosiven unter lateraler und nasaler Verschlußlösung [tl, pm, ...], Kombinationen mit entstimmten Konsonanten, nichtsilbische Vokale;
- Endhalbsilben und Rudimente (572): "klassische" Endhalbsilben ohne finale Obstruenten (außer [ç] und [x]) sowie einige nicht der Phonotaktik des Deutschen entsprechende Einheiten;
- Suffixe (88): Folgen stimmloser Obstruenten (ungerundet und gerundet);
- Konsonant-Konsonant-Diphone (167): Ausnahmeinventar zur Modellierung von Koartikulationseffekten an Silbengrenzen (Reduktion oder Assimilation); Elemente, die durch die Phonotaktik nicht erfaßt werden;
- Vokal-Vokal-Diphone (67): vor allem Übergänge von Vokalen zum Schwa [ə] und zu vokalisiertem /r/ [ɐ];
- "Neutrale Silben": (a) Silben mit Schwa, vor allem häufig auftretende Vor- und Nachsilben (75); (b) Silben mit silbischem Konsonanten [nach Elision eines Schwa] (122).

Insgesamt umfaßt dieses Inventar 2177 Bausteine und liegt damit im Rahmen üblicher datengesteuerter Systeme, die auf ein konventionelles Diphon- oder Halbsilbeninventar zurückgreifen. Wie Hörversuche gezeigt haben (Portele, 1996a), wurde in Präferenztests dieses Inventar deutlich besser bewertet als ein reines Diphon- oder Halbsilbeninventar.

Im Unterschied zu Diphon- oder Halbsilbeninventaren ergibt sich bei der Synthese hier nicht notwendigerweise eine lineare Abfolge der Elemente. Der Bausteinauswahl kommt bei der Synthese einer Äußerung somit eine wichtige Bedeutung zu. Um jeden Silbenkern wird zunächst unabhängig von den benachbarten Silbenkernen eine maximale Umgebung definiert, die unter Verwendung des Silbenkerns mit den Elementen des Inventars synthetisiert werden kann. Folgen stimmloser Frikative werden durch Suffixe realisiert. Setzt man hieraus die Äußerung zusammen, so werden sich an verschiedenen Stellen Überlappungen ergeben. Diese werden durch ein besonderes Regelwerk beseitigt; wichtiges Kriterium hierbei ist das Prinzip der maximalen Anfangskonsonantenfolge, das so viele Konsonaten wie möglich der jeweiligen Anfangshalbsilbe zuschlägt.

2. Akustische Synthese

Die Einteilung der Methoden akustischer Synthese in die zwei Kategorien *parametrische Synthese* (Abschnitt 2.1) und *Signalmanipulation* (Abschnitt 2.2) richtet sich nach der Art und Weise, in welcher das Anregungssignal behandelt wird. Bei rein parametrischer Synthese ist das

Anregungssignal stets künstlich; die Vokaltraktparameter können dagegen einem Regelsatz entstammen oder aus natürlichen Sprachdaten gewonnen sein. Systeme mit Signalmanipulation im Zeitbereich arbeiten stets mit Segmenten natürlicher Sprachsignale; halbparametrische Repräsentationen (beispielsweise Prädiktorkoeffizienten plus Residualsignal) sind hierin eingeschlossen.

2.1 Parametrische Synthese

Die meisten parametrischen Synthetisatoren folgen dem Quelle-Filter-Modell. Ein Signalgenerator (der im einfachsten Fall ein periodisches, impulsförmiges Signal für stimmhafte Laute, weißes Rauschen für stimmlose Laute erzeugt) liefert das Eingangssignal des linearen Filters, das die Übertragungsfunktion des Vokaltrakts modelliert. Zwischen Filter und Generator bestehen üblicherweise keine Rückkopplungen. Dieses Quelle-Filter-Modell wird bei parametrischer Sprachsignalübertragung (Vocoder) weithin benutzt. Das Ziel parametrischer Sprachübertragung ist jedoch Datenreduktion, die dort im Vergleich zur Sprachqualität (vor allem im Vergleich zur Natürlichkeit) immer Priorität besitzt. Hieraus ergibt sich ein Konflikt mit der Forderung der Sprachsynthese nach möglichst guter Sprachqualität.

Für die Datenreduktion in Sprachübertragungssystemen sind drei Kenngrößen maßgebend: 1) Parameterabtastrate, 2) Parameterquantisierung sowie 3) die Struktur des Synthetisators. Die ersten beiden Punkte betreffen ausschließlich die Datenübertragungsrate und sind für die Synthese nicht relevant. Die Frage der Struktur des Synthetisators ist auch eine qualitative und daher komplizierter. In parametrischen Synthesesystemen werden vorwiegend Formant- und Prädiktionssynthetisatoren eingesetzt. In Vocodern entscheidet man sich meist zugunsten der linearen Prädiktion, da diese robust und für die automatische Analyse gut geeignet ist und darüber hinaus einfache Quantisierungsschemata zuläßt. In TTS-Systemen wird häufig dem Formantsynthetisator der Vorzug gegeben, da ein Großteil des akustisch-phonetischen Wissens über die Eigenschaften von Lauten in Form von Formantfrequenzen und Formantübergängen vorliegt (Delattre, 1968), die sich wiederum leicht in Syntheseregeln umwandeln lassen. Die verhältnismäßig einfache Struktur des Synthetisators jedoch, wie sie in der parametrischen Sprachübertragung benutzt wird, ist für Sprachsynthesesysteme nicht geeignet.

Der größte Mangel sowohl von Prädiktions- als auch von Formantsynthetisatoren besteht darin, daß sie ein Allpolfilter, also ein reines Resonanzmodell darstellen; Laute, die Antiformanten enthalten, sowie die zugehörigen Übergänge werden damit nicht gut modelliert. Komplexer strukturierte Synthetisatoren schaffen hier Abhilfe (Klatt, 1980). Die größere Flexibilität solcher Synthetisatoren bringt andererseits eine wesentlich erhöhte Zahl einstellbarer (und einzustellender) Parameter mit sich, die es zunehmend erschweren, die zugehörigen Regeln zu formulieren. Beispielsweise verlangt der Synthetisator von KLATTALK (Klatt, 1982) 19 Parameter; der Synthetisator GLOVE des KTH-Systems (Carlson *et al.*, 1991) benötigt sogar 37.

Die Qualität des synthetischen Sprachsignals hängt auch vom Generator für das Anregungssignal ab. Einfache Systeme verwenden einen Rauschgenerator für stimmlose Laute, einen Impulsgenerator für stimmhafte Laute und einen Schalter, um zwischen den beiden Anregungsarten hin- und herzuschalten. Aufgefeiltere Systeme erlauben es, die beiden Generatoren nebeneinander zu betreiben und damit eine Art gemischter Anregung herzustellen. Neben der Frage der Interaktion zwischen dem stimmhaften und dem stimmlosen Anregungssignal existiert noch das Problem, die Stimmbandschwingung selbst zu modellieren. Ein reiner Impulsgenerator reicht nicht aus; er liefert eine ”summende” Qualität des Signals durch schlechte Modellierung der Feinstruktur der Stimme bei Frequenzen unterhalb 1500 Hz (Childers und Wu, 1990).

Häufig verwendet wird das vereinfachte Zwei-Massen-Modell der Stimmbänder nach Ishizaka und Flanagan (1972). Daneben ist auch das Modell von Liljencrants und Fant (LF-Modell; Fant *et al.*, 1985) im Einsatz; hier wird ein künstliches Anregungssignal, bestehend aus Versatzstücken von Kosinusschwingungen, direkt modelliert, wobei die spektralen Eigenschaften dieser künstlichen Anregungsfunktion durch entsprechende Parameter in weiten Grenzen gesteuert werden können. Dieses Modell wird auch im KTH-System (Carlson *et al.*, 1991) verwendet.

Die parametrische Synthese einer Frauenstimme stellt ein besonderes Problem dar (Klatt, 1987; Karlsson, 1989). Es genügt nicht, ausgehend von Daten über männliche Sprecher die Grundfrequenz anzuheben, den glottalen Öffnungsquotienten zu erhöhen und alle Formantfrequenzen mit einer Konstanten zu multiplizieren (Klatt, 1987). Zwischen männlichen und weiblichen Stimmen bestehen darüber hinaus noch mehr Unterschiede, beispielsweise in der Intonation (Möbius *et al.*, 1991). Durch die höhere Grundfrequenz besitzen Frauenstimmen überdies eine größere Interaktion zwischen der Grundfrequenz und dem ersten Formanten; es ist deswegen hier erheblich schwieriger, die Eigenschaften des Anregungssignals von denen des Vokaltrakts zu separieren.

Mehr auf Grundlagenforschung als auf praktische Anwendung hin orientiert ist die artikulatorische Sprachsynthese (Coker, 1976; Maeda, 1982; Schroeter und Sondhi, 1992; Kröger *et al.*, 1995); eine gute Übersicht über diese Verfahren mit zahlreichen Literaturhinweisen findet sich in (Kröger, 1996). Artikulatorisch basierte Sprachsynthese verwendet ein geometrisches Modell des Vokaltrakts; die Steuerung erfolgt direkt durch Simulation der Stellung und Bewegung der Artikulatoren. Hinsichtlich der Sprachqualität bleibt die artikulatorische Sprachsynthese hinter dem Quelle-Filter-Modell zurück (Kröger, 1996); wissenschaftliche

Bedeutung hat sie für die artikulatorische Phonetik. Das in Köln entwickelte Modell (Kröger *et al.*, 1995) verwendet die Artikulationsgeste als grundlegende Steuerungseinheit und erlaubt daher beispielsweise, Meßergebnisse und Hypothesen der artikulatorischen Phonetik durch Modellierung und (Re-)Synthese zu überprüfen.

Ein großer Mangel dieser Methode bestand bisher darin, daß direkte Messungen von Artikulationsvorgängen meist nur sehr ungenaue Ergebnisse lieferten oder – bei Verwendung der Röntgentechnik – wegen der damit verbundenen Strahlenbelastung für den Sprecher nur Daten in geringem Umfang abwarfen. In Form des Artikulographen ist jedoch jetzt eine verbesserte Meßtechnik verfügbar, durch die sich artikulatorische Daten in größerem Umfang sammeln lassen. Damit ist zu erwarten, daß die artikulatorische Sprachsynthese sich auf mehr Daten und damit realistischere Modelle als bisher stützen kann. Es erscheint sogar möglich, einen artikulatorischen Sprachsynthetisator direkt datengesteuert, beispielsweise mit Hilfe eines neuronalen Netzes, zu implementieren.

2.2 Synthese durch Signalmanipulation im Zeitbereich

In einem datengesteuerten System mit Verkettung vorgefertigter natürlichsprachlicher Elemente erstreckt sich der größte Teil aller Manipulationen auf die drei Parameter Amplitude, Dauer und Grundfrequenz. Die Manipulation spektraler Eigenschaften beschränkt sich auf die unmittelbare Nachbarschaft der Verkettungspunkte.

Die Idee, Sprachsignale direkt im Zeitbereich zu manipulieren, ist nicht neu (vgl. Rodet, 1977, 1980, oder Endres und Großmann, 1974). Es ist wohlbekannt, daß die Impulsantwort des Vokaltrakts sich annähernd aus gedämpften Schwingungen zusammensetzt, deren Frequenzen gleich den Formantfrequenzen und deren Dämpfungseigenschaften gleich den Bandbreiten der Formanten sind. Eine ungestörte Formantschwingung dauert jedoch länger als eine Grundperiode in fließender Sprache. Eine Grundperiode ist also nur ein Teil der Antwort des Vokaltrakts auf mehrere aufeinanderfolgende Anregungsimpulse. Nichtsdestoweniger können wir jede Grundperiode für sich als ein Signalsegment betrachten, das einen elementaren akustischen Baustein des Sprachsynthesesystems bildet. Derartige Bausteine lassen sich zu synthetischer Sprache zusammensetzen, und die drei prosodischen Parameter Grundfrequenz, Intensität und Dauer lassen sich ohne Übergang in den Frequenzbereich oder in eine parametrische Darstellung manipulieren. Da jede Grundperiode zudem einem Originalsignal entnommen werden kann, ist es prinzipiell möglich, die natürliche Sprachqualität zu erhalten.

Die Frage ist allerdings, wieviel an Qualität verlorengeht, wenn manipulierte Grundperioden zu einem Signal zusammengesetzt werden. Einfache Verkettung, ohne die Abtastwerte an den Übergangsstellen zu beachten, ergibt Sprünge, die die Qualität drastisch verschlechtern. Dies hat jahrelang den Einsatz dieser Methode verhindert, bis mit dem PSOLA-Verfahren (*"pitch-synchronous overlap add"*) ein Weg gefunden wurde, die "Bruchstelle" zwischen aufeinanderfolgenden Grundperioden über die ganze Periode zu verteilen und damit weitgehend unhörbar zu machen (Hamon *et al.*, 1989; Charpentier und Moulines, 1989). Auch PSOLA baut auf Grundperiodensegmenten als elementaren Bausteinen auf. Ein derartiger Baustein besteht hier aus einem Intervall von ungefähr zwei Grundperioden, die durch ein Fenster gewichtet werden, das um einen Anregungsimpuls zentriert ist. Damit überlappen sich benachbarte Bausteine in ihrer zeitlichen Abfolge, und das Ausgangssignal wird erzeugt, indem die Bausteine nach geeigneter Manipulation von Dauer, Grundfrequenz und Intensität aufaddiert werden.

Wird PSOLA direkt in Verbindung mit Sprachsignalen eingesetzt (PSOLA im Zeitbereich, TD-PSOLA), so ist es möglich, die Dauer, die Intensität und die Grundperiode zu ändern; spektrale Eigenschaften des Signals lassen sich dagegen nicht manipulieren. Dies wird mit zwei weiteren PSOLA-Algorithmen möglich: PSOLA im Frequenzbereich, eine verhältnismäßig komplexe Prozedur, sowie PSOLA in Verbindung mit linearer Prädiktion (LP-PSOLA). Bei LP-PSOLA wird das Signal einer Prädiktionsanalyse unterzogen; der Grad des Prädiktorfilters wird derart gewählt, daß die Information über die Grundfrequenz auf jeden Fall in das Residualsignal geht, wohingegen die Information über die Übertragungsfunktion des Vokaltrakts in den Filterkoeffizienten repräsentiert ist. Mit dieser Konfiguration können spektrale und temporale Eigenschaften des Signals separat modifiziert werden.

Solche Verfahren können sogar dafür eingesetzt werden, Übergänge durch Regeln zu modellieren, die im Inventar der Bausteine nicht enthalten sind. In einer älteren Version von HADIFIX (Portele *et al.*, 1994) war ein solches Verfahren implementiert, das die Modellierung von Vokal-Vokal-Übergängen durch Interpolation der LP-Spektren und grundperiodensynchrone Mischung der Residualsignale gestattete.

Dem PSOLA-Verfahren gebührt zweifelsfrei das Verdienst, die nichtparametrische Sprachsynthese überhaupt erst ermöglicht und damit eine erhebliche Verbesserung der Sprachqualität gegenüber den parametrischen Verfahren bewirkt zu haben. Nichtsdestoweniger hat das Verfahren einige Schwachpunkte. Während die Dauer des synthetischen Signals in weiten Grenzen variiert werden kann, wird bei Modifikation der Grundfrequenz mit TD-PSOLA die Qualität des synthetischen Signals entscheidend schlechter, wenn die Abweichung der Grundfrequenz gegenüber dem Original ungefähr eine halbe Oktave nach oben oder unten überschreitet. In einer halbparametrischen Repräsentation, wie sie beispielsweise bei LP-PSOLA eingesetzt wird,

kann dagegen auch die Grundfrequenz in weiteren Grenzen variiert werden.

Weiterhin beklagt wurde verschiedentlich die Tatsache, daß es bei der Modifikation der Grundfrequenz durch die überlappende Addition benachbarter Grundperioden zur Auslöschung eines einzelnen höheren Formanten kommen kann (Dutoit und Leich, 1993). Neben diesem Mangel sieht Dutoit die generelle Notwendigkeit, durch eine spektrale Normierung und die Herbeiführung einer strengen Monotonie der Bausteine des Inventars spektrale Unebenheiten an den Verkettungsstellen soweit wie möglich zu beseitigen. Ein ähnliches Verfahren ist auch in dem deutschen TTS-System PHRITTS (P. Meyer *et al.*, 1993) implementiert.Die so behandelten Bausteine liefern tatsächliche eine "glattere" synthetische Sprache; andererseits leidet dabei doch die Verständlichkeit. Bei aller Einfachheit des Verfahrens darf nicht übersehen werden, daß es sich bei PSOLA um eine nichtlineare Operation handelt, und daß somit die mehrfache Anwendung dieses Verfahrens, wie sie bei streng monotonen Bausteinen notwendig ist [einmal bei der Normierung der Grundfrequenz beim Erstellen der Bausteine, zum zweiten Mal dann bei der Synthese], zusätzliche nichtlineare Verzerrungen erzeugt.

Bei halbparametrischer Darstellung mit Hilfe der linearen Prädiktion kann unter gewissen Umständen auf PSOLA ganz verzichtet werden; dies liefert teilweise sogar bessere Ergebnisse (Macchi *et al.*, 1993). Bei LP-PSOLA wird das Residualsignal mit Hilfe des PSOLA-Algorithmus manipuliert; ist jedoch die Energie des Residualsignals auf wenige, nahe beieinanderliegende Abtastwerte konzentriert, so sind die übrigen Abtastwerte einer Grundperiode annähernd Null, und das Residualsignal kann ohne Qualitätsverlust hart abgeschnitten und neu zusammengesetzt oder – wenn eine Grundperiode verlängert wird – mit Nullwerten aufgefüllt werden.

2.3 Auf dem Weg zum "Personal Synthesizer" – Inventarerstellung und Sprecheradaption

Einer der großen Vorteile datengesteuerter Synthesesysteme ist der vergleichsweise geringe Aufwand bei der Erstellung einer neuen "Stimme". Die zugehörigen Arbeitsschritte sind 1) Aufsprechen der Trägersätze für das Inventar sowie 2) Gewinnen der Bausteine aus den Trägersätzen.

Die Einheiten sind in der Regel in kurze Trägersätze eingebettet; da die gleichen Einheiten für betonte und unbetonte Silben herangezogen werden, hat es sich bewährt, sie in den Trägersätzen in nebenbetonte Silben zu plazieren (Portele, 1996a).

Das Aufsprechen der Trägersätze erfordert bei der üblichen Zahl von Bausteinen (zwischen 1000 und 3000) eine Aufnahmesitzung von mehreren Stunden. Für den Sprecher bedeutet dies eine erhebliche Anstrengung, da die Trägersätze möglichst gleichmäßig ohne Variation von Lautstärke, Sprechgeschwindigkeit oder Stimmqualität gesprochen werden müssen. Von der Einhaltung der Sprechdisziplin hängt die Qualität des Syntheseinventars entscheidend ab. Nur wenige geübte Sprecher schaffen dies in einer Sitzung; meistens sind mehrere Sitzungen an aufeinanderfolgenden Tagen notwendig. In fast allen Sitzungen mißraten darüber hinaus einzelne Einheiten; dies stellt sich in der Regel erst beim Schneiden des Inventars heraus und erfordert weitere Termine mit dem Sprecher.

Aus der synthetischen Sprache ist der Originalsprecher gut wiederzuerkennen; nichtsdestoweniger bedeutet eine angenehme und klare Stimme sowie eine deutliche Aussprache noch nicht die Eignung einer Sprecherin oder eines Sprechers als Lieferant(in) eines Syntheseinventars. Die sprechereigene Prosodie wird nicht mit übernommen, und die Standardprosodie des Synthesesystems kann sich als unverträglich mit der Stimmqualität herausstellen. Außerdem klingt – aus welchen Gründen auch immer – die synthetische Stimme zumeist "härter" als das Original. Es ist somit stets zweckmäßig, mit einem neuen Sprecher zunächst Probeaufnahmen zu machen und ein kleines Inventar zu erstellen; erst wenn dieses sich bewährt, soll das gesamte Inventar aufgesprochen werden.

Das Schneiden des Inventars kann nach wie vor nicht vollautomatisch durchgeführt werden. Automatische Grobsegmentierungsalgorithmen (Boëffard *et al.*, 1993; Wesenick und Schiel, 1995) erlauben es zwar, die Bausteine in den Trägersätzen zu lokalisieren und zeitlich einzugrenzen; obwohl einzelne dieser Verfahren schon recht gut sind (Traber, 1996 mit Hinweis auf Boëffard *et al.*, 1993), muß die Feinabstimmung doch von Hand gemacht werden. Ebenfalls grundsätzlich von Hand durchzuführen ist eine Annotierung der Bausteine, beispielsweise das Markieren von Lautgrenzen in silbenorientierten Bausteinen. Der Arbeitsaufwand für die Erstellung eines kompletten Inventars liegt bei Verwendung der rechnergestützten Hilfsmittel in der Größenordnung von einer Arbeitswoche. Damit ist es durchaus realistisch, datengesteuerte Synthesesysteme mit mehreren Stimmen zu erstellen.

Portele, Stöber *et al.* (1996) stellen ein Verfahren vor, mit dem ein bereits geschnittenes Inventar auf neue Inventare der gleichen Art abgebildet werden kann; dies ergibt somit eine weitgehend automatische Segmentierungsprozedur. Die Trägersätze für das neue Inventar werden zum einen von dem neuen Sprecher aufgesprochen, zum anderen vom System mit einem vorhandenen Inventar synthetisiert. Ein DTW-Algorithmus bildet die neuen Trägersätze auf die synthetischen Sätze ab. Hierbei werden nicht nur die Einheitengrenzen übertragen, sondern auch sonstige für die Synthese erforderliche Zeitmarken (z. B. Lautgrenzen innerhalb eines Bausteins). Das Verfahren wurde mit zwei vorhandenen Inventaren (von einer Sprecherin und einem Sprecher) sowie mehreren Parametrisierungsmethoden getestet; normierte und nicht normierte Mel-Cepstrum-Koeffizienten erwiesen sich als fast gleichwertig; Übereinstim-

mung des Geschlechts der Sprecher der aufeinander abgebildeten Inventare war nicht erforderlich. Das neue Inventar wurde letztendlich in einem konkurrierenden Ansatz auf vier Kombinationen von (vorhandenem) Inventar und Parametrisierung abgebildet, wobei durch eine strenge Zeitvorgabe untaugliche Segmentierungen ausgeschieden werden konnten. Im praktischen Einsatz war die Segmentierung zu 67% korrekt; weitere 8% der Segmentgrenzen waren 10 ms, weitere 12% weniger als 50 ms vom Optimum entfernt, und bei 13% ergaben sich größere Abweichungen. Der Ansatz ist ziemlich rechenaufwendig; eine Workstation SUN SPARC 10 war rund 51 Stunden mit der Segmentierung von 2200 Elementen beschäftigt. Die manuelle Durchsicht und Korrektur ging schneller: bei etwa 100 Einheiten je Stunde war das Inventar mit 22 Arbeitsstunden durchsegmentiert. Hinzu kommt noch der Arbeitsaufwand für das Aufsprechen.

Obwohl ein "Personal Synthesizer", der es einem Benutzer nach Belieben ermöglicht, z. B. seine eigene Stimme oder die einer vertrauten Person zur Grundlage des Sprachsynthesesystems zu machen, ein wenig wie Spielerei klingen mag, ist dieser Aspekt nicht zu unterschätzen. Zum einen liegt diese Entwicklung im allgemeinen Trend zu fortschreitender Individualisierung, zum anderen kann es aber z. B. bei der Anwendung im Behindertenbereich von Vorteil sein, wenn der Vorleseautomat für einen Blinden von der Stimme einer vertrauten Person abgeleitet ist. In jedem Fall jedoch ist es vorteilhaft, in einem Synthesesystem mehrere Stimmen zur Verfügung zu haben.

Laut Sorin (1994) wird in Frankreich und auch in anderen Ländern daran gearbeitet, die Flexibilität von Ansagediensten und Auskunftssystemen durch den Einsatz textgesteuerter Sprachsynthese zu erhöhen. Da TTS noch nicht die erforderliche Qualität aufweist, wird an einem hybriden Modell gearbeitet. In einem derartigen System existieren feste Ansagen, deren Text sich nicht ändert, und variable Anteile. Die festen Ansagen werden aus Qualitätsgründen als Ganzes von einem Sprecher aufgesprochen und über reproduktive Sprachsynthese ausgegeben. Wird für die variablen Ansagen TTS verwendet, so müssen diese nahtlos in das Gerüst der festen Ansagen eingepaßt werden. Hierzu ist es notwendig, daß die Sprecher(innen) der festen Ansagen auch noch für die Erstellung von Syntheseinventaren herangezogen werden. Mindestens zwei Stimmen werden gebraucht: eine für An- und Absage, die andere für die Nachricht selbst (Sorin, 1994:54).

Ein weiteres Einsatzgebiet für die Sprachsynthese ist "Speech-to-Speech-Translation", also die sprachliche Kommunikation zwischen zwei menschlichen Kommunikationspartnern verschiedener Muttersprache unter Zwischenschaltung eines maschinellen Übersetzungssystems mit akustischer Ein- und Ausgabe [vgl. VERBMOBIL (Wahlster, 1993)]. Eine ergonomische Anforderung an die beteiligte Sprachsynthese besteht darin, die Stimme des Originalsprechers und die synthetische Stimme für seine übersetzte Äußerung einander ähnlich zu machen, so daß der Sprecher aus der synthetischen Äußerung in der anderen Sprache wiedererkannt, zumindest aber von anderen Sprechern unterschieden werden kann.

Diese Aufgabe erfordert eine sprecheradaptive Synthese, da ein Inventar des Originalsprechers in der fremden Sprache nicht erstellt werden kann. Hierbei beschränkt sich die Adaption nicht nur auf wenige stationäre, grobe Merkmale, wie mittlere Grundfrequenz oder mittlere Sprechgeschwindigkeit, sondern soll auch feinere Eigenschaften, z. B. spektrale Eigenschaften des Sprechers, auf die Synthese übertragen.

Im Rahmen von VERBMOBIL wurde diese Aufgabe grundsätzlich angegangen (Kraft, 1995; Rinscheid, 1995, 1996), indem das entwickelte datengesteuerte Synthesesystem für das Deutsche um eine experimentelle Sprecheradaption erweitert wurde. Ausgangspunkt sind hierfür zunächst mehrere Syntheseinventare, die den Sprecherraum so gut wie möglich ausfüllen sollen. Für die Synthese wird dann die synthetische Stimme ausgewählt, die der Stimme des Originalsprechers am ähnlichsten ist; die Auswahl des Inventars erfolgt nach Geschlecht, spektraler Ähnlichkeit und Ähnlichkeit der mittleren Grundfrequenz. Inventarauswahl und Grobanpassung der prosodischen Merkmale bilden also die erste Stufe.

Im laufenden Betrieb wird zunächst die mittlere Sprechgeschwindigkeit ebenso wie die mittlere Grundfrequenz der synthetischen Stimme an die entsprechenden Daten des Originalsprechers angepaßt. Dem überlagert kann eine weitere Anpassung der spektralen Eigenschaften der Synthese an die Klangfarbe der Originalstimme erfolgen. Der Ansatz von Rinscheid (1995) verwendet eine Merkmalskarte nach dem Prinzip von Kohonen (1989). Diese trainierbare Karte bildet die stimmlichen Eigenschaften des Synthesesprechers auf die des Originalsprechers ab. Sind genügend Daten vorhanden, so können im Prinzip einzelne Grundperiodenfenster, wie sie für PSOLA verwendet werden, direkt in die Karte eingespeichert werden; hierbei ergibt sich so etwas wie eine Vektorquantisierung der Stimme, und die ausgewählten Grundperioden der Synthesestimme können bei der Synthese direkt durch die in der Merkmalskarte an gleicher Position befindlichen Signale des Originalsprechers ersetzt werden. Das Verfahren, das in der Spracherkennung durchaus vielversprechend ist, ist jedoch für die Sprachsynthese nicht brauchbar, da durch Diskontinuitäten im Zielsignal die Synthesequalität bis zur Unbrauchbarkeit verschlechtert wird. Als Ausweg bleibt eine zeitvariante, d.h., lautabhängige spektrale Adaption der Sprechereigenschaften mit einem zeitveränderlichen Filter. Dessen Implementierung jedoch ist zeitaufwendig; ein zeitinvariantes Filter läßt sich in Echtzeit einsetzen, liefert aber etwas schlechtere Adaptionsergebnisse.

2.4 Optimierung der Verkettung in datengesteuerten Systemen

Bei der Verkettung natürlichsprachlicher Einheiten besteht eines der Probleme darin, daß die Signale des Inventars bei verschiedenen Bausteinen nie ganz gleich sind, weil keine Äußerung des Sprechers exakt reproduzierbar ist. Somit entstehen an den Verkettungsstellen notwendigerweise Diskontinuitäten in den spektralen Eigenschaften des Signals, die die Qualität vermindern.

Die wirksamste Methode zur Vermeidung dieser Diskontinuitäten ist die sorgfältige Auswahl des Sprechers und das Einhalten der notwendigen Sprechdisziplin beim Aufsprechen der Bausteine. Diesem Anspruch hält jedoch nicht jeder Sprecher stand. Also müssen zusätzliche Maßnahmen ergriffen werden, die teils bei der Inventarerstellung, teils bei der laufenden Synthese zum Einsatz kommen. Verfahren wie das der harmonischen Analyse und Resynthese (P. Meyer *et al.*, 1993; Dutoit und Leich, 1993) sind in der Lage, eine spektrale Normierung und Angleichung an den Bausteingrenzen durchzuführen; wegen der damit verbundenen mehrfachen Anwendung von PSOLA wird die Qualität aber im ganzen verschlechtert.

Zur Optimierung der Verkettungsstelle in laufender Synthese untersuchte Kraft (1994, 1995) die Frage, inwieweit eine leichte Verschiebung der Verkettungsstelle entlang der Zeitachse die Verkettung optimieren kann. Da die Signale der einzelnen Bausteine einander stets ein Stück weit überlappen, ist eine Verkettungsstelle nicht von vornherein festgelegt, sondern kann in einem bestimmten Zeitrahmen (z. B. 50 ms) variieren. Dies kann dazu ausgenutzt werden, den günstigsten Übergangspunkt zu finden, d. h., die genaue Verkettungsstelle dort anzusetzen, wo die spektrale Distanz zwischen den entsprechenden Stützstellen beider Bausteine am geringsten ist.

Will man das absolute Minimum der spektralen Distanz erreichen, so ergibt sich ein zweidimensionales Problem, dessen Komplexität quadratisch mit der Zahl der "verkettungsfähigen" Stützstellen der beiden Signale wächst; der Rechenaufwand wird so erheblich, daß der Einsatz dieser Verkettungsstrategie eine Synthese in Echtzeit nicht mehr erlaubt. Die Komplexität läßt sich auf ein lineares Maß reduzieren, wenn die Stützstellen der beiden Bausteine in der Umgebung der Verkettungsstelle einander fest zugeordnet werden und dann dort verkettet wird, wo bei dieser Auswahl die spektrale Distanz am geringsten ist.

Wie zudem Hörversuche ergeben haben (Kraft, 1994), ist die Verbesserung der Sprachqualität durch Einsatz dieser Strategie im Mittel nicht signifikant; in Einzelfällen läßt sich jedoch eine sehr deutliche Verbesserung erzielen. Damit liegt der Nutzen dieses Verfahrens nicht so sehr in einer Verbesserung der Qualität der laufenden Synthese, sondern darin, daß dieses Verfahren ein Werkzeug an die Hand liefert, das es gestattet, die Konsistenz eines Inventars und damit die Artikulationstreue des Sprechers zu evaluieren und bereits bei der Erstellung des Inventars eine Aussage über mögliche Problembausteine und die zu erwartende Güte der Verkettung zu gewinnen

3. Zur Frage der Prosodie

Neben der Qualität auf segmentaler Ebene ist eine gute Prosodiesteuerung entscheidend für die Qualität des synthetischen Sprachsignals. Hierbei wird die Qualität der Prosodie um so wichtiger, je besser die segmentale Verständlichkeit des Systems ist.

Der klassische Ansatz der Prosodie umfaßt die drei linguistischen Parameter Quantität (Dauer), Intensität (Betonung) und Intonation (Melodie), die sich vorwiegend auf die akustischen Parameter Dauer und Rhythmus sowie Grundfrequenz abbilden; demgegenüber spielt die Amplitude (Lautheit) als prosodischer Parameter auf akustischer Ebene eine geringere, wenn auch nicht zu vernachlässigende Rolle.

Die Prosodiesteuerung betrifft die symbolische und die akustische Ebene der Synthese gleichermaßen. Sie ist mit allen linguistischen Ebenen verknüpft, insbesondere auch mit Semantik und Pragmatik (Kohler, 1991). Dementsprechend wird die Prosodiesteuerung in einem TTS-System, das keine semantische oder pragmatische Analyse durchführen kann, immer Stückwerk bleiben: *"Even if one considers that the congruence between syntax and prosody is not complete, neither of these operations* [prosodic segmentation of the utterance and generation of adequate prosodic contours, WH] *can be carried out effectively (i.e., so as to mimic the naturalness of human speech) without the availability of an in-depth linguistic description of the utterance"* (Sorin, 1994:58). In TTS-Systemen heißt das, entweder den Text prosodisch zu annotieren oder sich mit einer Standard-Prosodie zu begnügen [die bei ausgefeilten Systemen allerdings schon recht gut ist (Kohler, 1991, 1996; Traber, 1992, 1996; vgl. auch Kraft und Portele, 1995)]. Bei inhaltsgesteuerter Sprachsynthese (vgl. Abschnitt 5.3), wo die semantische Information zugänglich ist, kann eine Verbesserung der Prosodie des synthetischen Signals erwartet werden, obwohl unser Wissen über die Umsetzung von Semantik und Pragmatik in die Prosodie des Sprachsignals auch bei prosodisch besser erforschten Sprachen, wie Englisch oder Deutsch, noch lückenhaft ist (Sorin, 1994).

Auf symbolischer Seite besteht die Aufgabe zunächst darin, einen Satz korrekt in prosodische Phrasen zu segmentieren und ggf. auch die Gewichtigkeit der Phrasengrenzen festzulegen. Die zweite große Aufgabe betrifft die Festlegung der Akzente. Wortakzent (Wortbetonung) einerseits sowie Satz- und Phrasenakzente sind streng auseinanderzuhalten (Kohler, 1991). Grundsätzlich kann jede Silbe, die Trägerin der Wortbetonung ist, auch den Satz- oder Phrasenakzent tragen, und letzterer wird (von Ausnahmen abgesehen, die fast ausschließlich auf Emphase und Kontrastbetonungen beschränkt sind) stets auf eine (wort-)betonte Silbe fallen. Jedes Wort enthält eine betonte Silbe, aber nicht jedes Wort im Satz erhält eine Betonung. Dies bedeutet, daß die

Zuweisungsregeln für Satz- und Phrasenakzente im wesentlichen Deakzentuierungsregeln sind, die die Akzente auf den (wort-)betonten Silben entweder abschwächen oder die Silben völlig deakzentuieren, wie dies beispielsweise für Funktionswörter zutrifft. Auf die weitere Darstellung dieser im wesentlichen (computer-)linguistischen Aufgabe muß aus Platzgründen verzichtet werden. Es sei nochmal bemerkt, daß diese Aufgabe nur unvollständig gelöst werden kann, wenn keine Information aus Semantik und Pragmatik zur Verfügung steht.

Bekanntermaßen reicht die Symbolverarbeitung eine Zeichenkette an die Verkettungsstufe weiter, die die zu synthetisierende Information in Lautschrift (ggf. parallel dazu auch orthographisch) enthält. Die Prosodiesteuerung reichert diese Zeichenkette mit prosodischen Steuerzeichen (für Akzente, Phrasengrenzen usw.) an, die ggf. durch manuell eingefügte Annotationen ergänzt werden. Die signalseitige Prosodiesteuerung extrahiert diese Steuerzeichen und setzt sie um in Werte für Dauer, Rhythmus und die Grundfrequenzkontur.

Hierzu existieren zahlreiche Verfahren. Dauer und Intonation (Grundfrequenz) werden meist getrennt voneinander modelliert, einige Ansätze behandeln beide Parameter gemeinsam. Wie auf der segmentalen Ebene stehen auch hier regel- und datengesteuerte Verfahren einander gegenüber. Über einige von ihnen soll im folgenden kurz berichtet werden.

3.1 Dauersteuerung

In der Dauersteuerung haben sich im wesentlichen drei Ansätze herauskristallisiert:(1) das klassische Regelmodell der Lautdauersteuerung, bei dem die Dauer jedes Lautes durch eine Grammatik sequentiell abzuarbeitender Regeln spezifiziert wird (z. B. Klatt, 1979; Kohler, 1988); (2) ein multiplikativ-additives Modell (van Santen, 1993, 1994) auf Lautbasis sowie (3) Modelle, die mit größeren, beispielsweise silbenorientierten Einheiten arbeiten (z. B. Campbell und Isard, 1991).

Van Santen (1993, 1994) schlägt aufgrund früherer Arbeiten und eigener statistischer Messungen ein Modell vor, das die Dauer jedes Lautes abhängig von wenigen Parametern (Kontext, Stellung innerhalb einer Silbe, Stellung der Silbe in der Frase, Grad der Akzentuierung) spezifiziert, die entweder einen additiven oder einen multiplikativen Beitrag leisten, so daß sich die Gesamtdauer als eine Summe von Produkten berechnet. Die Berechnung erfolgt für jeden Laut kontextabhängig aufgrund eines Entscheidungsbaumes (dieser trennt zuerst Vokale und Konsonanten, dann z. B. innerhalb der Konsonanten intervokalische und nicht intervokalische Konsonanten usw.); die Werte der Parameter wurden aufgrund statistischer Untersuchungen an einem größeren Korpus festgelegt.

Die Elastizitätshypothese von Campbell und Isard (1991) sieht die Silbe als die Einheit, auf die sich die Dauersteuerung auswirkt. Die Dauer der einzelnen Laute in einer Silbe hängt von deren ”Elastizität”, d. h. Kompressionsfähigkeit, ab. Stimmhafte Verschlußlaute mit zeitlich eng begrenzter Dauer der Verschlußpause sind weniger ”elastisch” als Frikative oder Vokale. Für jeden Laut wurde die ”Elastizität” experimentell aus einem Korpus als Varianz der gemessenen Lautdauern festgelegt. Für die Gesamtdauer der Silbe wiederum sindmehreren Faktoren maßgebend: (1) die Zahl der Phoneme in der Silbe; (2) die Art des Silbenkerns (gespannter oder ungespannter Vokal, Diphthong oder silbischer Konsonant); (3) Position der Silbe in der Phrase; (4) Akzentuierung der Silbe; sowie (5) ob die Silbe in einem Funktions- oder einem Inhaltswort steht. In ihrer ”harten” Form besagt die Elastizitätshypothese, daß die Dauer für die Silbe als Ganzes festgelegt ist; innerhalb der Silbe verteilt sie sich auf die einzelnen Laute entsprechend deren ”Elastizität”. In der ”weicheren” Form wird zwischen verschiedenen Silbentypen unterschieden, insbesondere werden äußerungsfinale Silben wegen der finalen Längung gesondert behandelt.

In einer Untersuchung von H. Meyer *et al.* (1995) mit dem Bonner System HADIFIX wurde die Elastizitätshypothese getestet, indem eine Dauersteuerung auf Lautebene einer Dauersteuerung auf Silbenebene in einem Präferenztest gegenübergestellt wurde. Hierbei wurde die silbenorientierte Dauersteuerung signifikant bevorzugt und dementsprechend in HADIFIX implementiert.

3.2 Intonationssteuerung

Für die Synthese des Deutschen haben die Intonationssteuerungen von Kohler (1991, 1996) und Traber (1992, 1996) die bisher besten Ergebnisse aufzuweisen. Diese Modelle sollen deshalb kurz vorgestellt werden.

Das Kieler Intonationsmodell KIM (Kohler, 1991, 1996) ist regelbasiert und streng phonologisch ausgerichtet. Das Regelwerk ist mit Hilfe einer generativen Grammatik formuliert. Es beschreibt die grundlegende globale Makroprosodie von Phrasen und Sätzen und schließt die Mikroprosodie mit ein. Berücksichtigt werden folgende Bereiche: (1) Wortakzent, (2) Satzakzent, (3) Intonation, (4) prosodische Phrasierung (Phrasengrenzen), (5) globale Sprechgeschwindigkeit und ihre Änderung, (6) Register und (7) Verzögerungsphänomene. Ziel des Modells ist es, die gesamte prosodische Vielfalt − auch der Spontansprache − durch eine sehr begrenzte Zahl von Kategorien und das generative Regelsystem zu modellieren. Das Modell hat eine symbolische und eine parametrische Komponente. Die prosodischen Kategorien werden symbolisch festgelegt und mit einem distinktiven Merkmalssystem klassifiziert. Auf die symbolischen Regeln folgen parametrische, die numerische Werte (insbesondere Grundfrequenz und Dauer) kontextsensitiv für die verschiedenen symbolischen Kategorien festlegen (Kohler, 1996:90).

Jedes satzakzentuierte Wort erhält Intonationsmerkmale, die entweder Gipfel- oder Talkonturen sind und im ersteren Fall entweder einen monoton fallenden oder fallend-steigenden Grundfrequenzverlauf haben; die Talkontur hat entweder leicht oder stark ansteigen-

de Grundfrequenz. Wichtigstes Kennzeichen eines Gipfels ist seine Position in bezug auf die zugehörige betonte Silbe (hier gekennzeichnet durch den Beginn des betonten Vokals); Kohler unterscheidet zwischen frühem, mittlerem und spätem Gipfel, die bei sonst gleicher Äußerung bedeutungsunterscheidend werden. Das Modell verwendet keine Deklination, vielmehr ist sequentieller Abstieg ("Downstep") implementiert, d.h., jeder F0-Gipfel liegt um einen bestimmten Wert tiefer als der vorhergehende, wenn kein Neuansatz erfolgt.

In der TTS-Implementierung werden zunächst in der linguistischen Vorverarbeitung die Notationen bereitgestellt, die dann die Generierung der prosodischen Marken für die jeweiligen Eingabeketten steuern können. Da die linguistische Vorverarbeitung in TTS nur die syntaktische Ebene erfaßt, müssen diese Marken von Hand eingegeben werden, wenn die gesamte Vielfalt des Modells ausgeschöpft werden soll. Das Modell enthält aber eine gute Voreinstellung, die dann wirksam wird, wenn die semantische und pragmatische Information nicht verfügbar ist. In der parametrischen Ebene ist ein Grundfrequenzgipfel durch zwei oder drei distinkte Stützstellen realisiert, zwischen denen der Grundfrequenzverlauf interpoliert wird; Entsprechendes gilt für Talkonturen.

Zur Erleichterung der manuellen Eingabe prosodischer Annotationen entwickelte Kohler (1991, 1994/96) eine eigene Annotationssprache, die es erlaubt, im Eingabetext jeden prosodischen Parameter einzeln festzulegen. Die Sprache ist leicht zu handhaben und gibt dem Benutzer ein Maximum an Freiheitsgraden.

Traber (1992, 1995, 1996) beschreibt eine datengesteuerte Prosodiegenerierung mittels automatischer Lernverfahren. Derartige Lernverfahren, beispielsweise mit neuronalen Netzen, erlauben in der Regel eine Generalisierung der trainierten Daten, also hier die Möglichkeit, auch für nicht gelernte Äußerungen und Äußerungsstrukturen sinnvolle Grundfrequenzverläufe zu erzeugen.

Als Datenbasis dient ein prosodisches Korpus von etwa 2000 Sätzen mit rund 24000 Wörtern. Für das Erlernen der Grundfrequenzverläufe und Dauerwerte wurden die den Sprachsignalen zugeordneten Akzentwerte und Phrasengrenzen benötigt. Da die automatische Zuordnung des verwendeten CNET-Synthesesystems (CNET, 1991) häufig nicht korrekt war, wurde die Zuordnung manuell erstellt. Unterschieden wurden Silben mit starkem und schwachem Grundfrequenzakzent, akzentuierte und nebenbetonte Silben ohne Grundfrequenzakzent sowie unakzentuierte Silben. An Grenzen wurden unterschieden einfache Silbengrenzen, Wortgrenze ohne Phrasengrenze sowie verschiedene Typen finaler und nichtfinaler Phrasengrenzen.

Die Grundfrequenzsteuerung wird durch ein rückgekoppeltes neuronales Netz vom Elman-Typ vorgenommen. Dieses Netz enthält typischerweise zwei verdeckte Schichten mit 20 bzw. 10 Knoten und fünf Rückführungen von der zweiten verdeckten Schicht zur Eingabeschicht. Der Grundfrequenzverlauf innerhalb der Silbe wird durch fünf Stützstellen beschrieben (am Silbenanfang, an Anfang, Mitte und Ende des Silbenkerns sowie am Silbenende); als Eingabe wurden verschiedene binäre Parameterkombinationen verwendet, darunter eine mit 52 Parametern, die für jede Berechnung der Grundfrequenzwerte auch Information aus bis zu 6 benachbarten Silben holt. Das rekurrente Netz ist nach Traber (1996) in der Lage, deklinations- bzw. downstep-ähnliche Strukturen in der Intonationskontur nachzubilden.

Um Vereinfachung der symbolischen Beschreibung bemüht sich das Modell von Heuft und Portele (1996) im Rahmen von HADIFIX, indem es die gesamte Information über Akzentuierungen auf einen einzigen Parameter *Prominenz* abbildet. Im Unterschied zu Akzentuierungen, die meist mit wenigen kategorialen Stufen (typisch: nicht betont – nebenbetont – hauptbetont – verstärkt betont) beschrieben werden, werden der Prominenz hier Werte zwischen 0 und 31 zugewiesen. Diese sind damit keine Kategorien im linguistischen Sinne mehr, sondern Quantisierungsstufen eines quasikontinuierlichen und damit der akustischen Domäne zuzuordnenden Parameters. Dieser erhält seine Berechtigung durch vorausgegangene Perzeptionsexperimente, die zeigten, daß Hörer durchaus in der Lage sind, zwischen zahlreichen Prominenzstufen zu diskriminieren, und daß zwischen der wahrgenommenen Prominenz und beispielsweise der Silbendauer ein beinahe linearer Zusammenhang besteht. Ähnliche Resultate waren schon vorher von Fant und Kruckenberg (1989) sowie Terken (1991) berichtet worden. Auch für Phrasengrenzen wurde ein entsprechender Prominenzfaktor eingeführt, der zwischen 0 und 9 rangiert. Weiterhin werden mit diesem Modell Dauer und Grundfrequenz in einem integrierten Ansatz generiert. Das zugehörige Regelsystem ist noch in der Entwicklung, ebenso wie die Realisierung mit einem neuronalen Netz (Portele, Reuter, Heuft, 1996), die aber bereits vielversprechende Ergebnisse liefert.

4. Evaluierung der Qualität

Die Evaluierung der Performanz sprachverarbeitender Systeme hat in den letzten Jahren als eigenes Forschungsgebiet erhebliche Bedeutung erlangt. Einen guten Überblick über Methoden und Ergebnisse bieten Pols (1991, 1992, 1994) sowie Jekosch (1993). Ausgehend vom Ziel einer bestimmten Untersuchung ist nach Pols zu unterscheiden zwischen einer *globalen* und einer *diagnostischen* Evaluierung (wobei der Übergang zwischen beiden Methoden fließend ist). In der Sprachsynthese beurteilt die globale Evaluierung ein System (oder mehrere vergleichend) als Ganzes nach globalen Kriterien (Verständlichkeit, Natürlichkeit, Deutlichkeit der Aussprache usw.), während die diagnostische Evaluierung dazu dient, systematische Fehler aufzuspüren und zu lokalisieren. Sie liefert dem Systementwickler Kriterien an die Hand, sein System gezielt zu verbessern.

Hierfür ist es notwendig, den Beitrag jedes einzelnen Moduls so genau wie möglich einzugrenzen.

Sprachqualität ist eine mehrdimensionale Größe mit zahlreichen Aspekten. Dementsprechend zahlreich sind die Methoden zur Ermittlung qualitativer und quantitativer Aussagen. Da keine zuverlässige signalbasierte[2] Meßprozedur besteht, muß die Sprachqualität bzw. müssen die Teilaspekte auditiv,[3] d.h., in Hörversuchen, ermittelt werden. Hierbei wird von der Methodik unterschieden zwischen 1) Kategorien, die quantitativ gemessen werden können, vor allem Verständlichkeit und Verstehbarkeit, sowie 2) Kategorien, die nur global und qualitativ bestimmt werden können, z.B. Natürlichkeit oder Grad der Anstrengung beim Zuhören. Solche Größen werden üblicherweise in Einschätzungstests durch Skalierung ermittelt.

Für die Akzeptanz eines Sprachsynthesesystems sind die beiden Aspekte der Verständlichkeit und Verstehbarkeit einerseits und der Ermüdungseffekt andererseits die wichtigsten Größen. Einige Methoden ihrer auditiven Ermittlung sollen im folgenden kurz dargestellt werden.

4.1 Evaluierung der Verständlichkeit und Verstehbarkeit

Die *Verständlichkeit* (engl. *intelligibility*) wird auf mehreren Ebenen gemessen: auf der segmentalen Ebene (Laut- bzw. Silbenverständlichkeit), der Wortebene oder der Satzebene. Hierbei wird festgestellt, welcher Anteil der dargebotenen Elemente (Laute bzw. Silben im sprachlichen Zusammenhang; Wörter, Sätze) richtig erkannt, d.h. beispielsweise, korrekt nachgesprochen oder schriftlich korrekt wiedergegeben werden kann.

Unter *Verstehbarkeit* (engl. *comprehension*) soll dagegen ein Maß definiert sein, das es erlaubt, den Anteil eines zusammenhängenden Textes zu ermitteln, den die Versuchsperson dem Sinn nach verstanden hat. Die Versuchsperson wird hier beispielsweise über den Inhalt eines Textes befragt oder gebeten, den Text im Diktat niederzuschreiben.

4.1.1 Laut- und Silbenverständlichkeit (*"intelligibility"*). Zu ihrer Ermittlung existieren seit langer Zeit wohlbekannte und gut erprobte Testverfahren. Diese unterscheiden sich vor allem in der Form der Antwort der Versuchspersonen (Fellbaum, 1984):

- Bei der *offenen Antwortform* wird die Versuchsperson gebeten, die dargebotenen Stimuli schriftlich oder lautschriftlich so wiederzugeben, wie sie sie gehört hat. Die bekannten Logatom-Tests (Fellbaum, 1984) und insbesondere auch der CLID-Test (*Cl*uster *Id*entification; Jekosch, 1989, 1992, s.u.) bedienen sich dieser Antwortform.
- Bei der *geschlossenen Antwortform* erhält die Versuchsperson zu jedem dargebotenen Stimulus eine Liste von Alternativen zur Auswahl, von denen sie diejenige ankreuzen soll, die dem Höreindruck entspricht. Bei den gebräuchlichsten Tests, die wiederum mit Einsilbern arbeiten, unterscheiden sich die Alternativen in nur einem Laut bzw. einer Lautkombination, je nach Test wortinitial, -medial oder -final; deswegen sind diese Tests auch als *Reimtests* bekannt. Im Deutschen wird für Zwecke der Evaluierung von Sprachübertragungskanälen meistens der Reimtest von Sotscheck (1982) verwendet.

Das Hauptproblem bei der Entwicklung dieser Tests besteht darin, den für die jeweilige Anwendung wesentlichen Eigenschaften der untersuchten Sprache möglichst gerecht zu werden. Der Reimtest beispielsweise ist so gut wie möglich phonemisch ausbalanciert, d.h., die Lauthäufigkeit der Wörter des Tests entspricht der Lauthäufigkeit fortlaufender Texte des Deutschen. Dies ist für den Test von Sprachübertragungskanälen ein wesentlicher Gesichtspunkt, obwohl auch andere Aspekte, wie z.B. die Silbenstruktur des Deutschen, nicht vernachlässigt werden sollten (Sendlmeier, 1991).

In unserem Fall der Sprachsynthese, insbesondere wenn der Test zu diagnostischen Zwecken durchgeführt wird, ist die Frage der *Abdeckung* wesentlich, also die Frage, wie mit dem Test eine möglichst große Zahl von Elementarbausteinen bzw. Regeln des Synthesesystems erfaßt werden kann. Ist eine Regel ungeschickt formuliert oder ein Baustein schlecht artikuliert bzw. geschnitten, so ergibt das im späteren Betrieb des Synthesesystems einen systematischen Fehler, der sich stets dann auswirkt, wenn die fragliche Regel bzw. dieser Baustein zum Einsatz gelangt. Nicht zuletzt aus diesem Grund ist für Synthesesysteme ein Verständlichkeitstest vorzuziehen, der auf der Basis logatomähnlicher Bausteine und sinnleerer Wörter und − vor allem − mit offener Antwortform arbeitet (Pols, 1992; Jekosch, 1994). Nur so wird die Versuchsperson unbeeinflußt von der eventuellen Vorgabe einer Auswahl von Alternativen oder dem Vorhandensein bzw. Nichtvorhandensein lautlich ähnlicher sinnvoller Wörter den Höreindruck so niederschreiben, daß es für die diagnostische Auswertung eines Systems sinnvoll wird, also daß beispielsweise die Ergebnisse in Lautverwechslungsmatrizen zusammengefaßt werden können.

Zu diesem Zweck wurde in Bochum der Cluster-Identifikationstest (CLID-Test; Jekosch, 1989, 1992) entwickelt. Hierzu gehört ein programmierbarer Stimulusgenerator, der einsilbige sinnleere Wörter in phonetischer Transkription sowie orthographienaher Darstellung generiert, sowie ein Auswertungsmodul. Dieses transkribiert die Antworten der Versuchspersonen, ermittelt, nach Lautclustern getrennt, den Anteil der korrekten Antworten und stellt, wenn gewünscht, Verwechslungsmatrizen auf. Die Stimuli richten sich nach den phonotaktischen Gesetzmäßigkeiten des Deutschen (vgl. Kohler, 1977) sowie zusätzlichen einschrän-

[3] Entsprechend einem Vorschlag von Blauert (1994) sollen hier anstelle der Bezeichnungen *subjektiv* und *objektiv* [als Übersetzungen des englischen *subjective* bzw. *objective*] die von der Sache her besser zutreffenden Bezeichnungen *auditiv* bzw. *signalbasiert* verwendet werden.

kenden Bedingungen, die versuchsspezifisch festgelegt werden können (wenn es beispielsweise darum geht, gewisse Konsonantenfolgen bevorzugt zu testen). Der CLID-Test verwendet eine offene Antwortform. Wegen seiner Flexibilität ist er für Anwendungen in der Sprachsynthese besonders geeignet.

Neben der Laut- und Silbenverständlichkeit sind Wort- und Satzverständlichkeit wichtige Kenngrößen eines Sprachübertragungs- bzw. Sprachsynthesesystems. Sie sind mehr zur globalen Charakterisierung eines Synthesesystems geeignet als zur Ermittlung diagnostischer Information. Ihre Ermittlung stellt wegen des Lerneffekts (jeder Text kann pro Versuchsperson nur einmal verwendet werden!) hohe Anforderungen an den Umfang und die Erstellung des Stimulusmaterials. Auf die Darstellung der zugehörigen Testverfahren wird aus Platzgründen verzichtet.

4.1.2 Verstehbarkeit (*"comprehension"*). Bei der Verstehbarkeit geht es um das Verständnis eines zusammenhängenden, sinnvollen Textes und somit um die Fähigkeit des Systems zur Übermittlung des Inhalts einer Nachricht. Es ist wohlbekannt, daß zum fehlerfreien Verständnis eines Textes nicht jedes Wort und jeder Laut vollständig erkannt werden muß; vieles kann aus dem Kontext − nicht zuletzt auch über die Prosodie − ergänzt und erschlossen werden. Dies rechtfertigt die Definition und Ermittlung einer eigenständigen Größe *Verstehbarkeit* als Attribut für die Beschreibung von Sprachsynthesesystemen. Im Gegensatz zur Verständlichkeit, die in der Regel quantitativ definiert und ermittelt wird, läßt sich die Verstehbarkeit sowohl qualitativ als auch quantitativ definieren. Als Stimulusmaterial dienen in allen Fällen kurze, zusammenhängende, allgemeinverständliche Texte möglichst gleichen Schwierigkeitsgrades, die auch die lautliche Struktur (Lauthäufigkeit, Worthäufigkeit, Silbenstruktur usw.) der Sprache berücksichtigen sollen. Unter diesem Gesichtspunkt wurden z.B. von Sendlmeier und Holzmann (1991) eine Reihe von Textpassagen (zu je etwa 100 Wörtern) aus Radiosendungen zusammengestellt. Wird die Verstehbarkeit qualitativ ermittelt, so ist sie eines der Attribute einer skalierenden Bewertung. Ansonsten wird sie nach Anhören des Textes durch Befragung der Versuchsperson zum Inhalt des Textes oder durch Diktat des Textes ermittelt. Wie jedoch bereits die Versuche von Sendlmeier und Holzmann (1991) ergaben [und dies wurde durch die Evaluierung des Bonner Sprachsynthesesystems HADIFIX (Portele *et al.*, 1992) bestätigt], sind die Versuchspersonen recht zuverlässig in der Lage, ihr Verständnis eines gehörten Textes selbst einzuschätzen, so daß für diagnostische Zwecke ggf. auf die aufwendige Befragungs- oder Diktatmethode verzichtet werden kann.

4.2 Bewertung der Natürlichkeit und zugehöriger Attribute

Natürlichkeit bedeutet nicht notwendigerweise, daß die synthetische Stimme wie ein Mensch klingen muß; es mag vielmehr sogar wünschenswert sein, daß ein Zuhörer jederzeit in der Lage ist, die maschinelle Provenienz eines synthetischen Signals als solche zu erkennen. Andererseits darf es nicht anstrengender sein, einer synthetischen Stimme zuzuhören, als einer natürlichen.

Die Natürlichkeit einer synthetischen sprachlichen Äußerung kann global oder über eine Reihe von Detailfragen ermittelt werden (Sendlmeier, 1991; CCITT, 1987; ITU-TSS, 1993). Hierbei gelangt der Präferenztest ebenso zum Einsatz wie die auditive Bewertung anhand von Bewertungsskalen. Beim Präferenztest vergleicht die Versuchsperson je zwei Stimuli global miteinander und gibt an, welchem System sie den Vorzug geben würde. Bei der auditiven skalierenden Bewertung hört die Versuchsperson einen Stimulus (vorzugsweise eine mehrsätzige Textpassage) und bewertet dann das System anhand verschiedener Attribute, zu denen jeweils eine Bewertungsskala angegeben ist (beispielsweise eine Fünf-Punkte-Bewertung von *sehr gut* bis *sehr schlecht*). Diese Art der Bewertung ist in einer Vorlage des ITU-TSS (1993; früher CCITT) für Sprachübertragungsstrecken genormt und als hauptsächliche Bewertungsart vorgeschrieben. Wie sich jedoch gezeigt hat (Hess *et al.*, 1994; Kraft und Portele, 1995), ist dieses Verfahren allein, d.h., ohne flankierende quantitative Messung der Verständlichkeit oder Verstehbarkeit, jedoch für die detaillierte Bewertung von Sprachsynthesesystemen − auch zu nichtdiagnostischen Zwecken − nicht ausreichend.

4.3 Einige Studien im einzelnen

4.3.1 Vergleichende Bewertung verschiedener Systeme (Klaus *et al.*, 1993; Fellbaum *et al.*, 1994). Diese Studie, die im Jahr 1992 durchgeführt wurde, erfaßte 13 Sprachsynthesesysteme verschiedener Entwicklungsstufen, davon zwei regelbasierte Systeme mit Formantsynthetisator ohne Prosodiesteuerung, ein System mit artikulatorischer Synthese sowie vier datengesteuerte Systeme mit natürlichsprachlichen Bausteinen und Zeitbereichssynthese mit PSOLA; die restlichen Systeme waren TTS-Systeme mit Formantsynthetisator. Zur Kontrolle wurden vier Codierungsverfahren mitbewertet, die als Referenzpunkte dienten. Die Bewertung umfaßte zwei Schritte: (1) eine globale Einstufung mit Mean-Opinion-Score und (2) einen Diktattest zur Ermittlung der Verständlichkeit und Verstehbarkeit (Sendlmeier und Holzmann, 1991).

Der globale Einschätzungstest umfaßte die Einzelparameter Höranstrengung, Schwierigkeiten des Verstehens, Deutlichkeit der Aussprache, Aussprachefehler,

betonungsfehler, Sprechgeschwindigkeit, Annehmlichkeit der Sprache und Gesamtqualität; alle Attribute waren auf einer Fünf-Punkte-Skala zu bewerten.

Die Ergebnisse beider Tests bestätigten übereinstimmend die Überlegenheit der Zeitbereichssynthese, was die Verständlichkeit betrifft. Auch in der Gesamtbeurteilung lagern diese Systeme vorn.

4.3.2 Vergleichende diagnostische Evaluierung von fünf Synthesesystemen für das Deutsche im Rahmen von VERBMOBIL (Kraft und Portele, 1995; Hess *et al.*, 1994) *und daran anschließende Untersuchungen*. Das Verbundprojekt VERBMOBIL (Wahlster, 1993) hat als Ziel die Entwicklung eines Übersetzungssystems mit Ein- und Ausgabe in gesprochener Sprache. In diesem Rahmen ist auch ein Sprachsynthesesystem für das Deutsche mit bestmöglicher Sprachqualität zu entwikkeln, wobei die Optimierung der segmentalen Qualität und damit der Verständlichkeit im Vordergrund stand.

Eine zu Beginn der Arbeiten durchgeführte diagnostiche Evaluierung erfaßte fünf Synthesesysteme für das Deutsche. Ausgewählt wurden hierbei Systeme, die bei den beteiligten Projektpartnern aus dem Vorfeld des Projekts oder aus früheren Arbeiten zur Verfügung standen und als Ausgangspunkt für die Weiterentwicklung eines VERBMOBIL-Synthesesystems dienen konnten. Drei der Systeme (im folgenden als D1, D2 und D3 bezeichnet) verwendeten eine Zeitbereichssynthese mit natürlichsprachlichen Bauelementen, wobei eines der Systeme (D3) bereits auf besonders gute Segmentverständlichkeit optimiert war; die übrigen beiden Systeme (R1 und R2) waren regelgesteuert mit einem parametrischen Synthetisator auf Formantbasis, eines der Systeme (R1) besaß eine besonders ausgefeilte Prosodiesteuerung.

Die Untersuchung umfaßte drei Einzelbewertungen. Mit dem CLID-Test wurde die segmentale Verständlichkeit jedes Systems ermittelt. Weiterhin wurden die Systeme global jeweils paarweise in einem Präferenztest verglichen. Schließlich wurden die Systeme einzeln einer globalen, skalierenden Beurteilung anhand von 8 Attributen unterzogen. Die Versuchspersonen waren normalhörende Student(inn)en verschiedener Fakultäten der beiden beteiligten Universitäten ohne vorherige Erfahrung im Umgang mit Sprachsynthesesystemen.

Ermittlung der segmentalen Verständlichkeit mit dem CLID-Test. Die Versuchsergebnisse bestätigten die Erwartungen: an der Spitze lag das System D3, gefolgt von D2 und D1; die beiden parametrischen Synthesesysteme schnitten schlechter ab. Dies bestätigte die Ergebnisse früherer Untersuchungen (Fellbaum *et al.*, 1994;vgl. auch Sorin, 1994), daß eine rein parametrische Synthese stets mit einem Verlust an Verständlichkeit besonders bei den Konsonanten einhergeht.

Da der CLID-Test – ebenso wie der Reimtest – nur einsilbige Stimuli verwendet, kann die Performanz der Verkettung bei inneren Konsonantenfolgen (also zwischen zwei Silbenkernen) mit diesem Test noch nicht beurteilt werden. Hierzu wird es notwendig sein, den Test so zu erweitern, daß er auch zweisilbige Stimuli erzeugt, mit denen die inneren Konsonantenfolgen dann synthetisiert werden können.

Vergleichende Gesamtbeurteilung im Präferenztest. In diesem Test sollte die Satzverständlichkeit (in qualitativer Bewertung) und die Gesamtakzeptanz der Systeme verglichen werden. Der Test wurde als Präferenztest im Paarvergleich anhand von sechs Aussagesätzen durchgeführt. Jede Version eines Satzes wurde mit jeder anderen verglichen, wobei beide möglichen Reihenfolgen (A-B und B-A) je Satz und Systempaar einmal auftraten. Bei 5 Systemen und 6 Sätzen ergab dies insgesamt 120 zu bewertende Satzpaare. Das Präferenzkriterium wurde den 20 Versuchspersonen wie folgt erläutert: "*Sie sollen von den beiden Versionen diejenige auswählen, die Ihnen verständlicher erscheint. Ziehen Sie jedoch auch die Vorstellung in Betracht, Sie müßten täglich eine oder mehrere Stunden mit den Stimmen arbeiten.*"

Das regelgesteuerte System R1 erzielte das beste Ergebnis, gefolgt von den datengesteuerten Systemen D3 und D2; die Systeme R2 und D1 schnitten am ungünstigsten ab. Bei der Befragung gefiel den Versuchspersonen an den Systemen R1 und R2 eine gute Intonation und eine "glatte Artikulation," während die datengesteuerten Systeme vielfach als "deutlicher" und "klarer" bewertet wurden. Generell bescheinigten die Versuchspersonen der synthetischen Sprache eine hohe Verständlichkeit, stellten aber gleichzeitig heraus, daß der wesentliche Faktor für die Akzeptanz die Prosodie gewesen sei. Die Testergebnisse bestätigen diese Aussagen, indem das System R1 (mit der besonders ausgefeilten Prosodiesteuerung) das beste und das System D1 (ohne Prosodiesteuerung) das schlechteste Ergebnis erzielte. Der Test war somit kein Verständlichkeitstest im engeren Sinn, sondern ein globaler Akzeptanztest.

Globale Einzelbeurteilung anhand mehrerer Bewertungskriterien. Dieser Test sollte eine differenzierte Bewertung der Stärken und Schwächen einzelner Systeme bei der Erzeugung längerer Textpassagen in synthetischer Sprache ermöglichen. Zu diesem Zweck wurden die insgesamt 44 Versuchspersonen gebeten, jeden von einem System generierten Text durch acht Attribute zu bewerten. Alle Texte entstammen der *Liste der Passagen fließender Rede für die Sprachgüebeurteilung* (Sendlmeier und Holzmann, 1991). Eine als "bekannt" vorgegebene Passage, von allen Systemen synthetisiert, wurde zusätzlich von einer natürlichen (weiblichen) Stimme gesprochen und allen Versuchsteilnehmern vor Versuchsbeginn vorgespielt. Weitere fünf Texte wurden (nach Auslosung) je einem System fest zugeordnet und waren den Versuchspersonen vor Versuchsbeginn nicht bekannt. Jedes System wurde somit einmal mit dem bekannten, einmal mit einem unbekannten Text bewertet. Zur Eingewöhnung wurden den Versuchspersonen vor Beginn des eigentlichen Versuchs außer der Textpassage mit natürlicher Stimme die beiden ersten Testsätze (aus dem

vorangegangenen Präferenztest) in allen synthetischen Versionen vorgespielt.

Die verwendeten Attribute sowie deren Skalierung wurden der Empfehlung des CCITT (1987) entnommen. Im einzelnen waren zu bewerten (Frage zur Erläuterung in Klammern):

- *Verständlichkeit* (ist es leicht oder schwer, einzelne Laute/Wörter zu verstehen?)
- *Verstehbarkeit* (ist es leicht oder schwer, die Aussage des Textes zu verstehen?)
- *Deutlichkeit* (ist die Aussprache [Artikulation] eher klar oder undeutlich?)
- *Natürlichkeit* (klingt die gehörte Stimme eher natürlich oder unnatürlich?)
- *Annehmlichkeit* (ist der Klang der gehörten Stimme angenehm oder unangenehm?)

Diese 5 Attribute waren jeweils auf einer Sechs-Punkte-Skala zu bewerten (von *sehr schlecht* bis *sehr gut*, also z.B. von *sehr unnatürlich* bis *sehr natürlich*). Weiterhin waren zu bewerten

- *Aussprache* (wie störend finden Sie ggf. eine falsche Aussprache?) auf einer Fünf-Punkte-Skala von *sehr störend* bis *nicht störend*;
- *Betonung* (wie störend empfinden Sie ggf. eine falsche Bestonung?), Skala wie vorstehend;
- *Sprechgeschwindigkeit* auf einer Sechs-Punkte-Skala von *viel zu schnell* bis *viel zu langsam*.

Für die Frage der Akzeptanz von Sprachsynthesesystemen allgemein ist die globale Aussage dieses Tests von Bedeutung, weist sie doch darauf hin, wo nach Ansicht der Versuchspersonen die Stärken und die Schwächen der Systeme liegen. Das Ergebnis zeigt sehr klar, daß die schlechtesten Bewertungen bei der Beurteilung des Hörkomforts vergeben wurden. Drei Systeme wurden im Median als *ziemlich unnatürlich* beurteilt, eines als *unnatürlich* und das fünfte sogar als *sehr unnatürlich*; das Attribut *Annehmlichkeit* kam nicht besser weg: vier Systeme waren im Median *ziemlich unangenehm*, das fünfte *unangenehm*. Bezüglich Verstehbarkeit, Verständlichkeit und Deutlichkeit fielen die Beurteilungen um ein bis zwei Grade besser aus. Aussprache- und Betonungsfehler wurden als *ziemlich störend* bis *störend* empfunden.

Wie eine abschließende Faktorenanalyse der Ergebnisse zeigte, läßt sich ein großer Teil der Varianz der Bewertung nach Attributen durch zwei Faktoren erklären. Der erste Faktor umfaßt die Attribute *Verständlichkeit, Verstehbarkeit* sowie *Deutlichkeit*; der zweite die Attribute *Betonung* sowie (mit Anteilen auch von Faktor 1) *Natürlichkeit* und *Annehmlichkeit*; das Attribut *Aussprache* ist beiden Faktoren zuzuordnen. Damit wird der durch die Attribute aufgespannte Raum auf eine im wesentlichen artikulationsorientierte sowie auf eine im wesentlichen prosodieorientierte Dimension reduziert.

Die Entwicklung des Mischinventars (Portele, 1994; vgl. Abschnitt 1.3) ist bei dieser Untersuchung noch nicht berücksichtigt. Wie eine Gegenüberstellung synthetischer Stimuli mit diesem Inventar und natürlicher Sprache (Portele, 1996a) ergab, ist nunmehr die Lautverständlichkeit für Konsonanten im Anlaut und Auslaut für die synthetischen und natürlichen Stimuli fast gleich (zwischen 5 und 6% Fehler im Anlaut; zwischen 2 und 2% Fehler im Auslaut). Nur im Inlaut sind die synthetischen Stimuli mit 7% noch erheblich schlechter als die natürlichsprachlichen (2%). Dies geht im wesentlichen auf das Konto eines Lautes, des intervokalischen [h], bei dem noch ein systematischer Fehler im Inventar bzw. bei der akustischen Synthese zu verzeichnen war.

In einer weiteren Untersuchung (Jekosch *et al.*, 1995) wurde das in VERBMOBIL in der Entwicklung befindliche Synthesesystem *Sprechmobil* evaluiert. Die Evaluation umfaßte eine Messung der Verständlichkeit sowie den Einsatz der Synthese in einer Dialogsituation. Methodisch von Interesse ist vor allem die Bewertung des Systems in der Dialogsituation. Dialog heißt hier in erster Linie: Synthese und Anwender sprechen abwechselnd. Da das simulierte Dialogsystem unabhängig von der Leistung des Erkenners oder anderer VERBMOBIL-Module arbeiten sollte, wurden die Dialoge so formuliert, daß die Antworten des Benutzers auf den Verlauf des Dialogs keinen Einfluß hatten. Der Benutzer sollte jedoch seine Aufmerksamkeit ausschließlich auf die sprachliche Realisierung der synthetischen Stimme lenken; die Inhalte, Fragen und Aufgaben durften also die Versuchspersonen nicht überfordern.

Der Test wurde wie folgt realisiert. Die übliche Einführung in den Versuch wurde vom Synthetisator gesprochen; somit entstand ein längerer Monolog, diesem folgte ein Interview mit dem Probanden, bei dem einige leicht zu beantwortende Fragen zur Person und zur Situation dieses Tests gestellt wurden. Im zweiten Teil des Dialogs wurde dann in Anlehnung an das aus dem Fernsehen bekannte Quiz *Dingda* ein unterhaltendes Quiz veranstaltet, in dem die Probanden einige Begriffe erraten mußten. [Beispiel: *Da waren schon mal Menschen. Den sieht man manchmal abends. Er ist kleiner als die Erde.* (Mond)] – Somit hatte die Synthese verschiedene dialogische Aufgaben zu erfüllen: Aufforderungen mit der Bitte um persönliche Daten, Aussagesätze und Monologe bei Einführungen sowie quasi-spontansprachliche Formulierungen und Satz- bzw. Ja/Nein-Fragen.

Die anschließende Bewertung erfolgte durch zwei skalierte Beurteilungen. Zunächst wurden für die wesentlichen in CCITT (1987) genannten Attribute Annehmlichkeit, Artikulation (Verstehbarkeit), Aussprache, Betonung, Deutlichkeit, Sprechgeschwinsigkeit, Lautstärke, Natürlichkeit und Verständlichkeit "Schulnoten" verteilt; bis auf die Natürlichkeit, die die Note "mangelhaft" erhielt, lauteten die Noten auf "befriedigend" oder besser. Zur Kontrolle wurden die Versuchspersonen auch noch gebeten, eine Reihe gegensätzlicher Attribute (Portele, 1992) auf einer 11-Punkte-Skala zu bewerten. Die Stimmen wurden hierbei einerseits als sehr verständlich, deutlich und angenehm, andererseits aber als unmelodisch, eintönig, langsam, sehr unnatürlich und unpersönlich charakterisiert. Es bleibt

nachzutragen, daß *Sprechmobil* aufgabengemäß zuerst in Hinblick auf segmentale Qualität und Verständlichkeit optimiert wurde und die für die Natürlichkeit entscheidende Prosodiesteuerung erst an zweiter Stelle der Untersuchungen steht.

4.4 Zusammenfassende Diskussion

Zu den vorgestellten Untersuchungen ist bezüglich der untersuchten Synthesesysteme zusammenfassend festzuhalten:

- Datengesteuerte Systeme mit Zeitbereichssynthese haben Vorteile bei der Verständlichkeit auf segmenteller Ebene durch die bessere Realisierung von Konsonantenfolgen. Die Verständlichkeit (ebenso wie die Verstehbarkeit) der Sprachsynthese wird systemübergreifend bereits als ziemlich gut bis sehr gut beurteilt.
- Der Hörkomfort (vertreten durch die Attribute *Natürlichkeit* und *Annehmlichkeit* läßt noch erheblich zu wünschen übrig. Alle bewerteten Systeme bekamen hier mangelhafte Noten. Die Akzeptanz eines Sprachsynthesesystems als Ganzes jedoch steht und fällt mit der Natürlichkeit des akustischen Signals. Ist die Natürlichkeit gering (und damit die Stimme "verfremdet"), muß der Zuhörer angestrengt hinhören und ermüdet schnell.
- Ist die Verständlichkeit eines Systems gut, so wird bei der Gesamtbeurteilung einer guten Prosodiesteuerung hoher Stellenwert eingeräumt.
- Mit einem sorgfältig zusammengestellten Bausteininventar mit akustisch-phonetisch orientierten Einheiten gelingt es, synthetische Sprache zu erzeugen, die bezüglich der Lautverständlichkeit natürlichen Äußerungen nicht mehr nachsteht.

Zur Methodologie der Untersuchungen sei zusammenfassend bemerkt:

- Der CLID-Test hat sich bewährt, insbesondere lieferte er durch die offene Antwortform eine Fülle an diagnostischer Information in Form detaillierter Verwechslungsmatrizen. Eine Erweiterung auf zweisilbige Stimuli ist für künftige Entwicklungen wünschenswert.
- Die differenzierende Beurteilung der Systeme durch Attribute erlaubt es, die Gesamtakzeptanz von Sprachsynthesesystemen gezielt zu beurteilen. Die Attribute selbst werden offensichtlich zum einen einer vorwiegend artikulationsbezogenen, zum anderen einer vorwiegend prosodiebezogenen Dimension zugeordnet.

5. Einige ausgewählte Anwendungen

5.1 Einsatz der Sprachsynthese im Behindertenbereich

Als Hilfsmittel für Behinderte haben Sprachsynthesesysteme bisher ihr größtes Einsatzgebiet gefunden (Klatt, 1987; Fellbaum, 1996). Neben dem klassischen Einsatzgebiet als Vorleseautomaten für Blinde können Sprachsynthesesysteme auch für sprech- und hörbehinderte Personen gewinnbringend eingesetzt werden. Die folgende Aufzählung (nach Fellbaum, 1996) enthält hierzu einige wichtige Beispiele.

- Für Blinde: Vorleseautomat (Vorlesesystem), Textverarbeitungssystem mit Sprachausgabe, PC-Anwendungen, Warn- und Alarmsysteme.
- Für Sprechbehinderte: Umsetzung von eingegebenem Text in Sprache, Übersetzung von unverständlicher in verständliche Sprache.
- Für Taubstumme: Umsetzung von eingegebenem Text in Sprache.

In der häufigsten Anwendung, dem Vorleseautomaten für Blinde, wird die Sprachsynthese heute meistens mit einem optischen Scanner und der zugehörigen optischen Zeichenerkennungssoftware zu einem Gesamtsystem integriert. Hierbei ergeben sich eine Vielzahl praktischer Probleme, die häufig mehr mit der Ergonomie der Benutzeroberfläche zu tun haben als mit der Sprachsynthese. Einige Probleme jedoch haben direkt mit der Sprachsynthese zu tun und sollen anhand (Fellbaum, 1996) sowie anhand des Erfahrungsberichts von Portele und Krämer (1996) kurz vorgestellt werden.

Blinde sind nach einigem Üben in der Lage, synthetische Sprache mit sehr hoher Sprechgeschwindigkeit [bis 300 Wörter pro Minute (Klatt, 1987)] zu verstehen. Auf der anderen Seite wird auch eine stark verlangsamte Sprachausgabe verlangt. Das System muß also über eine Einstellmöglichkeit für die Sprechgeschwindigkeit in einem weiten Bereich (etwa 1:10) verfügen; hierunter darf jedoch die Verständlichkeit der Ausgabe nicht leiden. Es ist wohlbekannt, daß in natürlicher Sprache eine Änderung der Sprechgeschwindigkeit sich nicht auf alle Signalabschnitte gleichermaßen auswirkt; vielmehr sind dynamische Segmente, also Lautübergänge, die eine starke Bewegung der Artikulatoren erfordern, nur sehr begrenzt kompressionsfähig. Portele und Krämer (1996) haben deswegen einen zusätzlichen Gewichtungsfaktor für Segmente eingeführt, in denen sich die akustischen Eigenschaften der Bauelemente sehr schnell ändern und die deswegen nur wenig komprimiert werden dürfen.

Wenn der Benutzer ein Wort nicht versteht oder eine Passage nochmal hören will, wird er versuchen, das System anzuhalten und die fragliche Passage erneut aufzu-

rufen. Das System synthetisiert üblicherweise die Äußerungen Satz für Satz und gibt jeweils das Signal an die akustische Ausgabe weiter, die einen Satz im Hintergrund abspielt, während das System bereits mit dem nächsten beschäftigt ist. Kommt nun ein Interrupt vom Benutzer, so muß das System genau wissen, welches Wort gerade abgespielt wird, damit es an diese Stelle zurückspringen kann. Da nach einem nicht verstandenen Wort die Reaktion des Benutzers zudem noch meist verzögert eintrifft, besitzen manche Systeme eine Möglichkeit, sich nach einem Interruptvon Wort zu Wort zurückzuhangeln, bis der Benutzer das System wieder freigibt (Fellbaum, 1996).

Die Zeichenerkennungssoftware des optischen Lesegerätes ist meistens alles andere als perfekt. Besondere Schwierigkeiten hat sie beispielsweise mit Telefax-Vorlagen (Portele und Krämer, 1996). Wie Portele und Krämer berichten, ist das System HADIFIX im Anfangsstadium der Adaption an einen solchen Vorleseautomaten regelmäßig zusammengebrochen, wenn es ”Wörter” mit mehr als 300 Zeichen oder ”Sätze” mit mehr als 200 Wörtern zu synthetisieren hatte. Wenn die Lesesoftware darüber hinaus einzelne Zeichen nicht oder falsch erkennt, entstehen sinnleere Wörter und grammatikalisch unkorrekte Sätze. Solche ”Wörter” stehen in keinem Aussprachelexikon und sind auch für das implementierte Regelwerk zur Graphem-Phonem-Konversion in der Regel nicht zugänglich. Im Falls von HADIFIX besitzt die neben dem Aussprachelexikon eingesetzte regelbasierte Aussprachegenerierung (Stock, 1991) für jedes Zeichen einen Default-Lautwert, der dann angesteuert wird, wenn keine Regel mehr greift. Darüber hinaus ist das Bausteininventar so organisiert, daß auch die unsinnigste Lautfolge noch synthetisiert werden kann.

Ein Computerarbeitsplatz für Blinde ist stets auch mit einem Textverarbeitungssystem ausgestattet. Dieses muß an die Sprachausgabe angeschlossen sein, um dem Benutzer die Möglichkeit zu geben, über die Tastatur eingegebene Texte abzuhören und auf Schreibfehler zu überprüfen. Hierbei muß das System zum einen in einen Buchstabiermodus versetzt werden können (was bedeutet, daß Orthographie und Aussprache während der gesamten Verarbeitung nebeneinander verfügbar sein müssen), zum anderen soll eine Option existieren, die es erlaubt, die Satzzeichen mit auszugeben. In letzterem Fall spielt die prosodische Behandlung der Satzzeichen eine besondere Rolle. In HADIFIX werden sie an den vorangehenden Teil der Äußerung angeschlossen, aber prosodisch als eigene Phrase betrachtet (Portele und Krämer, 1996).

Ein nichttriviales Problem sind Eigennamen, Akronyme und Abkürzungen; sie können nicht für jeden Benutzer von vornherein festgelegt werden. In HADIFIX wird hierzu ein benutzerdefiniertes Zusatzlexikon zur Verfügung gestellt, in das der Benutzer diese Terme eingeben kann. Mit dem implementierten Regelwerk wird eine Aussprache generiert, die dann vom Benutzer nach Bedarf noch korrigiert werden kann.

Nach wie vor steht bei der Entwickling von Synthesesystemen im Behindertenbereich die Verständlichkeit der synthetischen Sprache an oberster Stelle. Wie jedoch die vorangegangenen Beispiele zeigen, sind Flexibilität und Robustheit zwei Aspekte, die im praktischen Einsatz und damit für die Akzeptanz des Systems ebenfalls eine entscheidende Rolle spielen.

5.2 Multilinguale Systeme

Multilingualität, wenigstens Zweisprachigkeit, ist eine sehr wünschenswerte Eigenschaft für ein Sprachsynthesesystem auch im Hinblick auf die Anwendungen (Sorin, 1994). So wird auch bei der Anwendung im Behindertenbereich immer wieder nach Systemen gefragt, die in mehreren Sprachen sprechen können (Fellbaum, 1996). Obwohl es bereits kommerziell erhältliche Systeme gibt, die in mehreren Sprachen sprechen, sind Vorleseautomaten für Blinde in den wenigsten Fällen derart ausgerichtet.

Eines der ersten multilingualen Systeme wurde von Carlson *et al.* (1982) vorgestellt; es bildete die Grundlage für das schwedische System Infovox. Weitere Systeme folgten u.a. bei CNET (Frankreich) (CNET, 1991; Bigorgne *et al.*, 1993), Lernout and Hauspie (Belgien) (1996) sowie bei AT&T (USA) (Sproat und Olive, 1994/96; Möbius *et al.*, 1996).

Als ein Beispiel sei die Architektur des AT&T-Systems ein wenig näher betrachtet. Das System besteht aus einer Kette von 13 Modulen, die streng in Reihe geschaltet sind: Textverarbeitung, Lemmatisierung, Akzentuierung und Prominenz, Aussprachegenerierung, Phrasierung, Phrasenakzente, Dauersteuerung, Intonationssteuerung, Steuerung der (Kurzzeit-)Amplitude, Anregungsfunktion, Auswahl der Bausteine, Verkettung und akustische Synthese. Die Schnittstellen zwischen den Modulen sind normiert, und jedes Modul fügt Information zum Datenstrom hinzu, bis schließlich die Verkettungsstufe die Daten in Parametersätze für die akustische Synthese umwandelt. Diese modulare Struktur mit streng normierten Schnittstellen hat u.a. den Vorteil, daß jedes Modul leicht herausgenommen, modifiziert und reintegriert werden kann; weiterhin ist es leicht möglich, an jeder Stelle diagnostisch in das System ”hineinzuhören”, um Ursachen für Fehler und Qualitätsverluste schnell zu finden.

Für die multilinguale Synthese (in diesem System neben Englisch neun Sprachen) ist es wichtig, daß die grundlegenden Algorithmen der Symbolverarbeitung wie auch der Verkettung und der akustischen Synthese sprachenunabhängig implementiert werden; sprachenspezifische Teile werden als Tabellen und Daten beigesteuert; somit kann quasi via ”Plug in” von einer Sprache zur anderen gewechselt werden.

Bei der Erweiterung des Systems auf eine neue Sprache muß zuerst die Phonotaktik untersucht und ein

Bausteininventar aufgebaut werden; Symbolverarbeitung und Prosodie folgen später, wobei die Grundversionen dieser Module meist schon eine halbwegs vernünftige Synthese ermöglichen, die als Arbeitsgrundlage für die weitere Entwicklung dient.

Bei rein regelgesteuerten Systemen, wie z. B. beim schwedischen System Infovox, kann die gleiche "Stimme" für alle Sprachen verwendet werden. Bei datengesteuerten Systemen ist dies nicht oder nur sehr bedingt möglich, wenn geeignete bilinguale Sprecher gefunden werden, die ein Inventar in zwei Sprachen in gleicher Qualität aufsprechen können. Sonst bedeutet bei diesen Systemen ein Wechsel der Sprache stets auch einen Wechsel der Stimme.

5.3 Inhaltsgesteuerte Sprachsynthese (Concept to Speech)

Als eine künftige Anwendung der Sprachsynthese zeichnet sich ihr Einsatz in Sprachdialogsystemen, Auskunftssystemen undautomatischen Ansagediensten ab (Sorin, 1994). Sofern nicht vorgefertigte Texte verwendet werden, wird die Sprachsynthese bei dieser Anwendung von einer Sprachgenerierungsstufe angesteuert. Deren Eingangsinformation besteht in einer abstrakten semantischen Repräsentation des auszugebenden Inhalts. Bei herkömmlicher Realisierung wandelt die Generierungsstufe diese in natürlichsprachlichen Text, der wiederum an die Sprachsynthese weitergereicht wird.

Diese Konfiguration bedeutet einen Umweg, der sich nachteilig auf die erreichbare Qualität des synthetischen Sprachsignals auswirkt. Die Generierungsstufe ist dahinggehend ausgelegt, in orthographischer Form einen Text zu erzeugen, der üblicherweise auf einem Bildschirm angezeigt wird. Ist anstelle des Bildschirms ein TTS-System nachgeschaltet, so muß dieses den Text wieder analysieren, Aussprache sowie Prosodie generieren und schließlich das Sprachsignal synthetisieren. Somit kann es den Vorteil nicht ausnutzen, daß die semantische Information am Eingang der Generierungsstufe vorhanden ist; vielmehr wirft die Generierungsstufe bei der Generierung der textlichen Repräsentation die semantische Information ab und gibt sie nicht an die Synthese weiter. Diesem Manko, das zu Lasten der Sprachqualität geht, kann eine inhaltsgesteuerte Sprachsynthese ("concept to speech", CTS) abhelfen.

Inhaltsgesteuerte Sprachsynthese als integrierter Bestandteil eines Sprachdialogsystems mit gesprochener Ausgabe ist andiskutiert (Sorin, 1994:59): *"... to have even the best TTS systems soundlike 'they know what they are talking about' will, still for a long time, be possible only in the case of their proper coupling with automatic text or message generators (as used in automatic man-machine dialogue systems)."* In der Verlängerungsphase für VERBMOBIL (Burgard *et al.*, 1996) ist ein solcher Ansatz in der Planung. Ein experimenteller CTS-Generator für ein Zugauskunfts-Szenario wurde in Nijmegen für das Niederländische entwickelt (Marsi, 1995, s.u.).

Worin liegt nun der Gewinn einer solchen inhaltsgesteuerten Sprachsynthese? Gewinnen können dabei beide Komponenten, die Sprachgenerierung wie auch die Synthese. Ein Vorleseautomat, der in der Regel keine oder höchstens eine sehr rudimentäre semantische Analyse des Textes durchführt (meist gerade genug, daß Aussprachefehler vermieden werden), "weiß" nicht, was er sagt. Infolgedessen lassen sich Phänomene wie Hervorhebungen, Kontrastbetonungen oder semantische Nuancen, die zu realisieren von einem geübten menschlichen Vorleser ohne weiteres verlangt werden kann, von einem Vorleseautomaten nicht erwarten. Auch wenn eine gut ausgefeilte Prosodiesteuerung eine natürlich klingende Standardprosodie zu erzeugen vermag, wird sie mit solchen Situationen nicht fertig. Da die Prosodie großenteils der Semantik und der Pragmatik verpflichtet ist (Sorin, 1994; Kohler, 1996), bleibt dieser Bereich dem Vorleseautomaten verschlossen.

Anders dagegen bei der inhaltsgesteuerten Sprachsynthese. Die für die Prosodie wichtigen Aspekte der semantischen Information z. B. Hervorhebungen, Kontrastbetonungen, fokussierte Satzteile und Satzakzente lassen sich aus der semantischen Repräsentation mitgenerieren, an den generierten Text anheften und an die Synthese weitergeben. Damit hat die Synthese dann alles, was sie braucht, um nicht nur eine von der Syntax abgeleitete Standardprosodie, zu erzeugen, sondern eine Prosodie, die die semantische Information unterstützt und sich deren Mittel zu bedienen weiß.

Jedoch nicht nur semantische Information ist bei der Generierung verfügbar, sondern auch syntaktische und lexikalische. Die Generierung erzeugt im Rahmen ihrer Grammatik korrekte Sätze und muß daher die gesamte syntaktische Information eines generierten Satzes ebenso verarbeiten wie die lexikalische und morphologische Information über Wortklassen oder Flexionen. Weitergereicht an die Sprachsynthese, kann dort die aufwendige und allem Fortschritt zum Trotz noch fehlerbehaftete syntaktische Analyse des Textes entfallen.

Werden Generierung und Synthese zu einer Stufe integriert, so kann auch die gesamte Graphem-Phonem-Konversion entfallen. Da jede Sprachgenerierungsstufe domänenorientiert arbeitet, wird ihr Lexikon stets kleiner sein als das eines Vorleseautomaten; zudem werden unbekannte Wörter nicht generiert. Benötigt wird im Lexikon der Generierungsstufe eine eigene Spalte, die Aussprache und Betonung jedes Wortes enthält; wortklassenabhängige Zusatzinformationen spezifizieren darüber hinaus beispielsweise, ob es sich um ein Funktionswort handelt, das im Satzgefüge deakzentuiert wird.

Der größte Teil der symbolverarbeitenden Module eines Vorleseautomaten kann folglich durch die inhaltsgesteuerte Sprachsynthese überbrückt und in die Generierung verlegt werden. Die Information, die die Generierung an die Synthese weitergibt, ist vollständiger (da Semantik und Pragmatik hinzukommen) und fehlerfreier als die Information, die sich das Sprachsynthese-

system aus dem generierten Text ohne die Zusatzinformationen selbst erschließen könnte.

Auch die Generierungsstufe kann in ihrer ureigensten Aufgabe der Wandlung einer Verbindung semantischer Konzepte in einen natürlichsprachlichen Text davon profitieren, daß gesprochene Sprache und nicht ein schriftlicher Text erzeugt wird. Erfolgt die Ausgabe in gesprochener Sprache, so kann die Generierungsstufe auf die gesamten Hilfsmittel zurückgreifen, die die Prosodie bereitstellt und die in geschriebenem Text nicht zur Verfügung stehen. Bei gesprochener Sprache hat der Sprecher im Unterschied zur Textform die Wahl, ob er eine Hervorhebung oder den Fokus einer Äußerung mit Hilfe der Prosodie oder mit Hilfe syntaktischer Konstruktionen, beispielsweise auf dem Weg über die Wortstellung realisiert. Die Realisierung mit Hilfe der Prosodie ist in der Regel die kürzeste und prägnanteste. Als ein einfaches Beispiel denke man daran, daß der Satz *"fahren Sie heute nach Stuttgart?*" fünf verschiedene Bedeutungen haben kann (und demnach auch fünf verschiedene Antworten erwarten läßt), je nachdem, ob der Satz neutral gesprochen wird oder eines der Wörter "fahren", "Sie", "heute" oder "Stuttgart" durch Kontrastbetonung hervorgehoben ist. Würden diese semantischen Unterschiede durch Veränderung der Wortstellung erreicht, so müßten fünf textlich verschiedene Sätze generiert werden, und mindestens vier davon wären länger als der vorstehend zitierte.

Macht die Integration von Generierung und Synthese damit syntheseseitig die symbolverarbeitenden Module überflüssig? Keineswegs, nur wird ihre Aufgabe eine andere sein. Als formale Repräsentation am Ausgang der Generierungsstufe ist annotierter Text bzw. annotierte Lautschrift zu erwarten. Die Annotationen, die sich von der Information über Wortklassen über die Syntax bis hin zur Semantik und Pragmatik erstrecken, werden syntheseseitig in prosodische Steuerinformation (z. B. Deakzentuierung von Funktionswörtern, Phrasierung, Einfügung von Pausen usw.) umgesetzt.

Der nachfolgenden akustischen Prosodiesteuerung steht über die semantische Information ein sehr viel reicheres prosodisches Inventar zur Verfügung als einem Vorleseautomaten. Die Prosodie einer inhaltsgesteuerten Sprachsynthese wird sich daher wesentlich stärker an spontansprachlichen Realisierungen orientieren können, als dies bei rein textgesteuerter Synthese möglich wäre. Zu erwarten ist also eine Steigerung der Sprachqualität vor allem im prosodischen Bereich.

Zum Zweck der Untersuchung der Interaktion zwischen der Generierung auf linguistischer Ebene und der Prosodie entwickelte Marsi (1995) eine Generierungsstufe für gesprochene Sprache im Rahmen eines Zugauskunftsszenarios. Das Modul erzeugt auf einem ersten Strang Text in orthographischer Darstellung und auf einem zweiten Strang eine prosodische Beschreibung in der ToBI-Notation (Silverman *et al.*, 1992). Das System generiert Auskünfte über Zugverbindungen zwischen zwei Bahnhöfen in den Niederlanden (mit Fahrplanangaben und eventuellen Umsteigepunkten). Das Modul Semantikkonstruktion erhält aus der Datenbasis von dem vorangeschalteten Informationsmodul die notwendigen Daten über den Reiseweg in Form konzeptionaler Konstituenten (Funktionen wie Zeit- oder Wegangaben; Instanzen wie Personen oder Dinge) und gruppiert diese zu einer semantischen Repräsentation. Insbesondere wird hier gekennzeichnet, welche Elemente neue Information enthalten (und daher per Arbeitsdefinition fokussiert sind). Das nachfolgende Syntaxmodul kleidet die semantische Repräsentation in eine gültige syntaktische Struktur. Soweit die Syntax dies zuläßt, steht die fokussierte Information stets am Satzende. Dem Syntaxmodul nachgeschaltet sind zwei Prosodiemodule zur Generierung von Akzent- und Phrasenstruktur und zur Generierung der Intonationskontur. Der Satzakzent wird festgelegt, indem alle nichtfokussierten Konstituenten deakzentuiert werden; innerhalb der fokussierten Konstituente(n) wird das Verb deakzentuiert, wenn ein weiteres akzentuiertes Wort vorhanden ist. Weitere Akzentuierungsregeln betreffen Komposita und Ausdrücke mit Zahlwörtern. Als nächstes wird die Phrasenstruktur festgelegt. Grundregel ist, eine Phrasengrenze hinter die oberste syntaktische Konstituente zu legen, die einen Akzent regiert, ohne den nächstfolgenden Akzent mit zu erfassen. Eine weitere Regel entfernt einen Teil der so gesetzten Phrasengrenzen wieder. Die Zuweisung der Gipfel und Täler der Intonationskontur erfolgt aufgrund der Phrasen- und Akzentstruktur. Die Umsetzung dieser formalen Darstellung in einen Grundfrequenzverlauf bleibt der Synthesestufe überlassen.

Dieses System ist ein Anfang: es dient dazu, die Interaktion von Semantik und Prosodie in einer Generierungsstufe zu erforschen. Inwieweit sich damit die Qualität der Synthese verbessern läßt, ist selbst noch Forschungsgegenstand.

Wichtig innerhalb der Generierungsstufe – und für CTS im allgemeinen – ist die Transparenz der Information. Wie Marsi (1995) betont, muß die Information über Fokus sowie semantische Funktionen und Instanzen an die Syntax und die Akzentuierungsstufe weitergegeben werden, um eine korrekte Deakzentuierung zu ermöglichen. In einem Übersetzungsprojekt wie VERBMOBIL (vgl. Burgard *et al.*, 1996) muß derartige Information noch von viel weiter her durchgeschleust werden, beispielsweise aus der Prosodie der Originaläußerung, wo die ursprünglichen Akzente und Phrasengrenzen lokalisiert werden, die dann in die Zieläußerung übertragen, in die andere Sprache transformiert und dort wieder für die Synthese verwendet werden. Darüber hinaus ist ein experimentelles sprachverarbeitendes System stets auch immer ein Werkzeug für die Grundlagenforschung (Kohler, 1991); es erlaubt, Hypothesen zu testen und die angewendeten Modelle zu verfeinern und zu optimieren. CTS könnte insbesondere dazu beitragen, den Durchgriff semantischer Katego-

rien und Konzepte auf die Prosodie näher zu untersuchen.

Eine praktische Möglichkeit, ein TTS-System alternativ auf CTS umzustellen, ist durch die Verwendung einer Annotationssprache gegeben. Eine solche, die im wesentlichen Prosodiezwecken dient, wurde bereits in Abschnitt 3.2 vorgestellt (Kohler, 1996). Portele (1996b) stellt einen weiteren Ansatz dieser Art vor, mit dem verschiedenste Steuerfunktionen ausgelöst werden können, bis hin zum Wechsel der Stimme. Im Hinblick auf Modularität und Portabilität verschiedener Sprachsynthesesysteme ist die sich abzeichnende Vielfalt individueller Entwicklungen solcher Annotationssprachen ungünstig (Portele, 1996b). Eine Möglichkeit, hier zu einem internationalen Standard zu kommen, wurde von Taylor und Isard (1995) mit SSML (*speech synthesis markup language*) aufgezeigt, einer universellen Annotationssprache, von der allerdings bis jetzt nur eine erste Demoversion vorliegt.

6. Schlußbemerkung

Eine Bemerkung bezüglich der Dichotomie "regelgesteuert vs. datengesteuert" erscheint an dieser Stelle angebracht. Sie betrifft für die Prosodie wie auch die segmentale Ebene. In diesem Beitrag wurden auf der segmentalen Ebene wie im Bereich der Prosodie regelgesteuerte und datengesteuerte Systeme einander gegenübergestellt. Die Implementierungen dieser beiden Systemtypen sind verschieden, aber in bezug auf das Wissen über die Sprache, das neben den Anwendungen stets ein weiteres Hauptziel aller Untersuchungen auch im Bereich der Sprachsynthese ist, sollen sich diese beiden Ansätze ergänzen und nicht gegenseitig ausschließen.Kein regelgesteuertes System kann heute mehr ohne umfangreiche experimentelle Untersuchungen an einem größeren Korpus erstellt werden (Kohler, 1991); ebensowenig kann ein datengesteuertes System ohne die Einbeziehung umfangreichen phonetischen und linguistischen Wissens mit Erfolg betrieben werden (Traber, 1996). Der Erfolg datengesteuerter Systeme auf der segmentaten Ebene liegt darin, daß natürlichsprachliche Bausteine direkt verwendet werden können und somit der Qualitätsverlust vermieden wird, der mit der rein parametrischen Repräsentation von Sprachsignalen verbunden ist. (In der Sprachcodierung hat die gleiche Tatsache dazu geführt, daß die parametrische Realisierung mit Vocoder durch nichtparametrische und halbparametrische Verfahren der direkten Signalcodierung weitgehend verdrängt wurde.) In der Prosodie stehen sich regelgesteuerte und datengesteuerte Verfahren eher als gleichwertige Partner gegenüber. Regelgesteuerte Systeme gehen von einer prosodisch-phonologischen Darstellung aus; die linguistischen Konzepte in der Symbolverarbeitung lassen sich daher unmittelbar umsetzen. Demgegenüber haben datengesteuerte Systeme den Vorteil, daß sie sich mit automatischen Lernverfahren trainieren lassen und somit mehr Daten in kürzerer Zeit verarbeiten können als der menschliche Systementwickler, der die Regeln aufstellt.

In diesem Beitrag wurden einige neue Entwicklungen in der Sprachsynthese aufgezeigt. Viele dieser Aspekte sind derzeit noch Forschungsgegenstand: inhaltsgesteuerte Synthese, die schnelle Entwicklung von Syntheseinventaren für zahlreiche Stimmen oder die sprecheradaptive Synthese. Verständlichkeit und Verstehbarkeit synthetischer Sprache sind heute schon mit den entsprechenden werten für natürliche Sprache vergleichbar. Vordringlichste Aufgabe ist und bleibt aber die Verbesserung der Natürlichkeit, denn hiervon hängt die Akzeptanz ab, die der Sprachsynthese für zahlreiche prospektive Anwendungen heute noch fehlt.

Bekanntmachung

Ein Teil der Arbeiten, über die hier berichtet wird, wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesminister für Bildung, Wissenschaft, Forschung und Technologie gefördert. Die Verantwortung für den Inhalt dieses Berichts liegt beim Verfasser.

Literatur

Allen, Jonathan (1992): "Overview of text-to-speech systems." In *Advances in speech signal processing;* ed. by M. M. Sondhi and S. Furui (Marcel Dekker, New York), 741-790

Allen, Jonathan / Hunnicutt, Sheri / Klatt, Dennis H. (1987): *From text to speech: The MITalk system* (Cambridge University Press, Cambridge)

Bailly, Gérard / Benoit, Christian (eds.) (1992): *Talking machines: theories, models, and designs* (North-Holland, Amsterdam)

Blauert, Jens (1994): persönliche Mitteilung an die ITG-Fachgruppe 6.4.3 (6.2.1994)

Bigorgne, D. / Boëffard, O. / Cherbonnel, B. / Emerard, F. / Larreur, D. / Le Saint-Milon, J. L. / Metayer I. / Sorin, C. / White, S. (1993): "Multilingual PSOLA text-to-speech system." In *Proc. IEEE ICASSP-93* (IEEE, New York), II-187-190

Boëffard, O. / Cherbonnel, B. / Emerard, F. / White, S. (1993): "Automatic segmentation and quality evaluation of speech units inventories for concatination-based multilingual PSOLA text-to-speech systeme." In *Proc. EUROSPEECH-93* (catalyst consult, Berlin), 1449-1452

Browman, Catherine P. (1980): "Rules for demisyllable synthesis using LINGUA, a language interpreter." In *Proc. IEEE ICASSP-80* (IEEE, New York), 561-564

Burgard, Christof / Karger, Reinhard / Wahlster, Wolfgang (Hrsg): *Wissenschaftliche Ziele und Netzpläne für Verbmobil Phase 2.* Verbmobil Technisches Dokument Nr. 44, Juli 1996 (DFKI, Saarbrücken)

Campbell, W. Nick / Isard, Stephen D. (1991): "Segment durations in a syllable frame." J. Phonetics *19*, 37-47

Carlson, Rolf / Granström, Björn / Hunnicutt, Sheri (1982): "A multi-language text-to-speech module." In *Proc. IEEE ICASSP-82* (IEEE, New York), 1604-1607

Carlson, Rolf / Granström, Björn / Karlsson, Inger (1991): "Experiments with voice modelling in speech synthesis." Speech Commun. *10*, 481-490

Carlson, Rolf / Granström, Björn / Nord, Lennart (1990): "Evaluation and development of the KTH text-to-speech system on the segmental level." Speech Commun. *9*, 271-277

Charpentier, Francis / Moulines, Eric (1989): "Pitch-synchronous waveform processing techniques for text-to-speech syn-

thesis using diphones." In *Proc. EUROSPEECH-89* (ENST, Paris), vol. 2, 13-19
Childers, Donald G. / Wu, Ke (1990): "Quality of speech produced by analysis-synthesis." Speech Commun. *9*, 97-117
CCITT (1987): Subjective quality assessment of synthetic speech (Contribution COM XII-176-E)
CNET (ed.) (1991): Note technique NT/LAA/TSS/430 – Recueil des publications et communications externes du département RCP (synthèse, reconnaissance de la parole et dialogue oral), janvier – décembre 1991 I: Synthèse à partir du texte et applications (CNET, F-22301 Lannion)
Coker, Cecil H. (1976): "A model or articulatory dynamics and control." Proc. IEEE *64*, 452-460
Delattre, Pierre (1968): "From acoustic cues to distinctive features." Phonetica *18*, 198-230
Dettweiler, Helmut (1984): *Automatische Sprachsynthese deutscher Wörter mit Hilfe von silbenorientierten Segmenten* (Diss., Tech. Univ., München)
Dettweiler, Helmut / Hess, Wolfgang J. (1985): "Concatenation rules for demisyllable speech synthesis." Acustica *57*, 268-283
Dutoit, Thierry / Leich, Henri (1992): "Improving the TD-PSOLA text-to-speech synthesizer with a specially designed multi-band excitation (MBE) re-synthesis of the segments database" In *Signal Processing VI*, Proc. EUSIPCO-92 (Brussels, Belgium), ed. by J. Vandewalle *et al.* (Elsevier, Amsterdam), 343-346
Endres, Werner K. (1973): "The transitional sounds of the German language as link elements for a speech synthesis." Acustica *26*, 33-36
Endres, Werner K. (1984): "Verfahren zur Sprachsynthese – ein geschichtlicher Überblick." Der Fernmeldeingenieur *38* (9), 1-37
Endres, Werner K.; Großmann, E. (1974): "Manipulation of the time functions of vowels for reducing the number of elements needed for speech synthesis." In *Proc. 1974 Speech Commun. Sem.* (2), 267 (Almqvist & Wiksell, Stockholm)
Endres, Werner K. / Wolf, Herbert E. (1980): "Speech synthesis for an unlimited vocabulary, a powerful tool for inquiry and information services" In *NATO ASI on Spoken Language Generation and Processing*, ed. by J. C. Simon (Reidel, Dordrecht), 411-428
Fant, Gunnar / Kruckenberg, Anita (1989): "Preliminaries to the study of Swedish prose reading and reading style." *STL-QPSR* (2), 1-83 (KTH Stockholm)
Fant, Gunnar / Liljencrants, Johan / Lin, Qi (1985): "A four-parameter model of glottal flow." *STL-QPSR* (4), 21-45 (KTH Stockholm)
Fellbaum, Klaus (1984): *Sprachsignalübertragung und Sprachsignalverarbeitung* (Springer, Berlin)
Fellbaum, Klaus (1996): "Einsatz der Sprachsynthese im Behindertenbereich." In *Fortschritte der Akustik, DAGA 96* (DEGA, Oldenburg), 78-81
Fellbaum, Klaus / Klaus, Harald / Sotscheck, Jochem (1994): "Hörversuche zur Beurteilung der Sprachqualität von Sprachsynthesesystemen für die deutsche Sprache." In *Fortschritte der Akustik – DAGA 94* (DEGA, Oldenburg), 117-122
Fujimura, Osamu (1976): Syllable as the unit of speech synthesis (Internal Report, AT&T Bell Laboratories, Murray Hill, NJ, USA)
Fujimura, Osamu / Lovins, J. B. (1978): "Syllables as concatenative phonetic units." In *Syllables and segments;* ed. by A. Bell and J. B. Hooper (North-Holland, New York), 107-120
Hamon, C. / Moulines, Eric / Charpentier, Francis (1989): "A diphone synthesis system based on time-domain modifications of speech." In *Proc. IEEE ICASSP-89* (IEEE, New York), 238-241
Hess, Wolfgang J. (1992a): "Speech synthesis – a solved problem?" In *Signal Processing VI*, Proc. EUSIPCO-92 (Brussels, Belgium), ed. by J. Vandewalle *et al.* (Elsevier, Amsterdam), 37-46
Hess, Wolfgang J. (1992b): "Sprachsynthese – ein gelöstes Problem?" In *Elektronische Sprachsignalverarbeitung* (3. gemeinsame Konf. der Techn. Univ. Berlin und Dresden und der Humboldt-Univ. zu Berlin), hg. von R. Hoffmann (Inst. f. Techn. Akustik, TU Dresden), 12-25
Hess, Wolfgang J. / Kraft, Volker / Portele, Thomas (1994): "Zum Problem der Evaluierung von Sprachsynthesesystemen – dargestellt am Beispiel der Synthesekomponenten in VERBMOBIL." In *Fortschritte der Akustik, DAGA 94* (DEGA, Oldenburg), 33-46
Heuft, Barbara / Portele, Thomas (1996): "Synthesizing prosody: a prominence-based approach." In *Proc. Intern. Conf. Spoken Language Processing* (ICSLP-96), Philadelphia, PA, USA
Ishizaka, K. / Flanagan, James L. (1972): "Synthesis of voiced sounds from a two-mass model of the vocal cords." Bell System Tech. J. *51*, 1233-1268
ITG-Fachgruppe 4.3.1 [Sotscheck, Jochem / Endres, Werner / Hess, Wolfgang / Hoffmann, Rüdiger / Krause, Manfred / Lacroix, Arild / Mangold, Helmut / Paulus, Erwin / Wolf, Herbert E.] (1996): ITG 4.3.1-01 Terminologie der Sprachakustik (unveröff. Manuskript)
ITU-TSS (1993): ITU-TSS draft recommendation P.8S: Subjective performance assessment of the quality of speech voice output systems (ITU Study Group 12 – Contribution 6 – COM 12-6-E)
Jekosch, Ute (1989): "The cluster-based rhyme test: a segmental synthesis test for open vocabulary." In *Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases,* Noordwijkerhout, The Netherlands, September 1989 (Institute of Phonetic Sciences, Amsterdam)
Jekosch, Ute (1992): "The cluster-identification test." In *Proc. Intern. Conf. on Spoken Language Processing (ICSLP-92)*, Banff, Alberta, Canada, 205-208
Jekosch, Ute (1993): "Speech quality assessment and evaluation." In *Proc. EUROSPEECH-93* (catalyst consult, Berlin), 1387-1394
Jekosch, Ute (1994): "Speech intelligibility testing: on the interpretation of results," J. Assessment of Voice I/O Systems (JAVIOS) *15*
Jekosch, Ute / Krause, Susanne / Mersdorf, Joachim (1995): *Evaluation der deutschsprachigen Synthese "Sprechmobil" im Verbmobilprojekt* (unveröff. Bericht, Universität Bochum)
Karlsson, Inger (1989): "A female voice for a text-to-speech system." In *Proc. EUROSPEECH-89* (ENST, Paris), 345-348
Klatt, Dennis H. (1979): "Synthesis by rule of segmental durations in English sentences." In *Frontiers of speech communication research*, ed. by B. Lindblom and S. Öhman (Academic Press, London), 287-300
Klatt, Dennis H. (1980): "Software for a cascade/parallel formant synthesizer." J. Acoust. Soc. Am. *67*, 971-980
Klatt, Dennis H. (1982): "The KLATTALK text-to-speech conversion system." In *Proc. IEEE ICASSP-82* (IEEE, New York), 1589-1592
Klatt, Dennis H. (1987): "Review of text-to-speech conversion for English." J. Acoust. Soc. Am. *82*, 737-793
Klaus, Harald / Klix, H. / Sotscheck, Jochem / Fellbaum, Klaus (1993): "An evaluation system for ascertaining the quality of synthetic speech based on subjective category rating tests." In *Proc. EUROSPEECH-93* (Catalyst Consult, Berlin), 1679-1682

Knohl, Lars / Rinscheid, Ansgar (1993), "Speaker normalization and adaptation based on feature-map projection", In *Proc. EUROSPEECH-93* (catalyst consult, Berlin), 367-370

Köster, Jens-P. (1973): *Historische Entwicklung von Syntheseapparaten zur Erzeugung statischer und vokalartiger Signale nebst Untersuchungen zur Synthese deutscher Vokale.* Hamburger Phonetische Beiträge, Bd.4. (Buske, Hamburg)

Kohler, Klaus J. (1977, ^{2}1995): *Einführung in die Phonetik des Deutschen* (Erich Schmidt, Berlin)

Kohler, Klaus J. (1988): "Zeitstrukturierung in der Sprachsynthese." In *Digitale Sprachverarbeitung*, ITG-Tagung, Bad Nauheim, hrsg. von A. Lacroix (VDE-Verlag, Berlin), 165-170

Kohler, Klaus J. (1990): "Segmental reduction in connected speech in German: phonological facts and phonetic explanations." In *Proc. of the NATO ASI on Speech Production and Speech Modeling,* Bonas, Gers, France, July 1989; ed. by W. J. Hardcastle and A. Marchal (Kluwer, Dordrecht), 69-92

Kohler, Klaus J. (1991): "Prosody in speech synthesis: the interplay between basic research and TTS application." J. Phonetics *19*, 121-138

Kohler, Klaus J. (1994/96): "Parametric control of prosodic variables by symbolic input in TTS synthesis." In *Proc. of the Second ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, NY, USA, 1994; to appear in *Progress in speech synthesis*, ed by J. Van Santen *et al.* (Springer, New York, Nov. 1996)

Kohler, Klaus J. (1996): "Modellgesteuerte Prosodiegenerierung: Die Implementation des Kieler Intonationsmodells (KIM) in der TTS-Synthese für das Deutsche." In *Fortschritte der Akustik, DAGA 96* (DEGA, Oldenburg), 90-91

Kohonen, T. (^{3}1989): *Self-Organization and associative memory"* (Springer, Berlin)

Kraft, Volker (1993): "Auditory detection of discontinuities in synthesis-by-concatenation." In *Proc. EUROSPEECH-93* (catalyst consult, Berlin), 929-932

Kraft, Volker (1994): "Does the resulting speech quality improvement make a sophisticated concatenation of time-domain synthesis units worthwhile?" In *Proc. of the Second ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, NY, USA, 65-68

Kraft, Volker (1995): *Konkatenation natürlichsprachlicher Bausteine zurSprachsynthese: Anforderungen, Methoden und Evaluierung* (Diss.,Ruhr-Universität Bochum)

Kraft, Volker / Portele, Thomas / Jekosch, Ute (1993): Primärevaluation und Bestandsaufnahme – VERBMOBIL-Teilprojekt 4 Sprachsynthese, Arbeitsschritt 4.1.1 (Unveröff. Bericht, Ruhr-Univ. Bochum)

Kraft, Volker / Portele, Thomas (1995): "Quality evaluation of five speech synthesis syetsm for German." Acta Acustica *3*, 351–366

Kröger, Bernd J. (1996): "Artikulatorische Sprachsynthese." In *Fortschritte der Akustik, DAGA 96* (DEGA, Oldenburg), 96-99

Kröger, Bernd J. / Opgen–Rhein, Claudia (1995): "A gesture-based dynamic model describing articulatory movement data." J. Acoust. Soc. Am. *98*, 1878-1889

Küpfmüller, Karl / Warns, O. (1956): "Sprachsynthese aus Lauten." Nachrichtentechnische Fachberichte *3*, 28-31

Kurzweil, Raymond (1976): "The Kurzweil Reading Machine: A technical overview." In *Science, Technology, and the Handicapped*, ed. by M. R. Redden and W. Schwandt (Amer. Assoc. for the Advancement of Science, Washington DC, USA, Report # 76R11), 3-11

Lernout and Hauspie (1996): TTS evaluator (L&H, Ieper, Belgium)

Lindblom, Björn E. F. (1963): "Spectrographic study of vowel reduction." J. Acoust. Soc. Am. *35*, 1773-1781

Macchi, Marian / Altom, Mary Jo / Kahn, D. / Singhal, S. / Spiegel, Murray (1993): "Intelligibility as a function of speech coding method for template-based speech synthesis." In *Proc. EUROSPEECH-93* (catalyst consult, Berlin)

Maeda, Shinji (1982): "A digital simulation method of the vocal-tract system." Speech Commun. *1*, 199-229

Marsi, Erwin (1995): "Intonation in a spoken language generator." In *Proceedings* (Dept. of Language and Speech, Univ. of Nijmegen), 85-98

Meyer, Horst / Portele, Thomas / Heuft, Barbara (1995): "Ein Silbendauermodell für die Sprachsynthese." In *Fortschritte der Akustik – DAGA 95* (DEGA, Oldenburg), 987-990

Meyer, P. / Rühl, H. W. / Krüger, R. / Kugler, M. / Vogten, L. L. M. / Dirksen, A. / Belhoula, K. (1993): "PHRITTS – a text-to-speech synthesizer for the German language." In *Proc. EUROSPEECH-93* (catalyst consult, Berlin), 877-880

Möbius, Bernd / Demenko, Grazyna / Pätzold, Mathias (1991): Parametrische Beschribung von Intonationskonturen. In *Beiträge zur angewandten und experimentellen Phonetik;* ed. by W. Hess und W. F. Sendlmeier. Beihefte zur Z. für Dialektologie und Linguistik, Bd. 72, 109-124 (Franz Steiner, Stuttgart)

Möbius, Bernd / Schroeter, Juergen / Van Santen, Jan / Sproat, Richard / Olive, Joseph (1996): "Recent advances in multilingual text-to-speech synthesis." In *Fortschritte der Akustik, DAGA 96* (DEGA, Oldenburg), 82-85

Moore, Roger K. (1994): "Twenty things we still don't know about speech." In *Progress and prospects of speech research and technology*, ed. by H. Niemann, R. De Mori, and G. Hanrieder (infix, St. Augustin), 9-17

Moulines, Eric / Charpentier, Francis (1990): "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones." Speech Commun. *9*, 453-467

O'Shaughnessy, Douglas (1987): *Speech communication. Human and machine* (Addison-Wesley, Reading, MA, USA)

Öhman, Sven E. G. (1966): "Coarticulation in VCV utterances: Spectrographic measurements." J. Acoust. Soc. Am., 151-168

Olive, Joseph P. (1990): "A new algorithm for a concatenative speech synthesis system using an augmented acoustic inventory of speech sounds." *In Proc. of the ESCA ETRW on Speech Synthesis*, Autrans, France (ESCA, Grenoble), 25-29

Peterson, Gordon E. / Sievertsen, E. (1960): "Objectives and techniques of speech synthesis" Language and Speech *3*, 84-95

Peterson, Gordon E. / Wang, W. / Sievertsen, E. (1958): "Segmentation techniques in speech synthesis." J. Acoust. Soc. Am. *30*, 739-742

Pols, Louis C. W. (1991): "Evaluating the performance of speech technology systems." Annual Report, Inst. of Phonetic Sciences, Amsterdam *15*, 27-42

Pols, Louis C. W. (1992): "Quality assessment of text-to-speech synthesis by rule" In *Advances in speech signal processing;* ed. by M. M. Sondhi and S. Furui (Marcel Dekker, New York), 387-418

Pols, Louis C. W. (1994): "Synthesis performance assessment." In *Progress and prospects of speech research and technology*, ed. by H. Niemann, R. De Mori, and G. Hanrieder (infix, St. Augustin), 63-68

Portele, Thomas (1993): "Evaluation der segmentalen Verständlichkeit des Sprachsynthesesystems HADIFIX mit der SAM-Testprozedur." In *Fortschritte der Akustik, DAGA-93* (DEGA, Oldenburg), 1032-1035

Portele, Thomas (1994): *Ein phonetisch-akustisch motiviertes Inventar zur Sprachsynthese deutscher Äuéerungen*. Diss., Univ. Bonn (Niemeyer,Tübingen, 1996)

Portele, Thomas (1996a): "Sprachsynthese durch Konkatenation natürlichsprachlicher Einheiten." In *Fortschritte der Akustik, DAGA 96* (DEGA, Oldenburg), 92-95

Portele, Thomas (1996b): "Annotationen, Tags und Kommandos in der Sprachsynthese." In *Fortschritte der Akustik, DAGA 96* (DEGA, Oldenburg)

Portele, Thomas / Höfer, Florian / Hess, Wolfgang (1994/96): "A mixed inventory structure for German concatenative synthesis." In *Proc. of the Second ESCA/IEEE Workshop on Speech Synthesis*, New Paltz,NY, USA, 1994; to appear in *Progress in speech synthesis*, ed by J. Van Santen *et al.* (Springer, New York, Nov. 1996)

Portele, Thomas / Krämer, Jürgen (1996): "Adapting a TTS system to a reading machine for the blind." In *Proc. Intern. Conf. Spoken Language Processing* (ICSLP-96), Philadelphia, PA, USA

Portele, Thomas / Reuter, André / Heuft, Barbara (1996): "Prosody generation with a neural network." In *Proc. 1996 World Congress on Neural Networks* (WCNN-96), San Diego, USA

Portele, Thomas / Sendlmeier, Walter F. / Hess, Wolfgang (1991): Das Sprachsynthesesystem HADIFIX. In *Beiträge zur angewandten und experimentellen Phonetik;* ed. by W. Hess und W. F. Sendlmeier. Z. für Dialektologie und Linguistik, Band 72, 143-154 (Franz Steiner, Stuttgart)

Portele, Thomas / Sendlmeier, Walter F. / Hess, Wolfgang / Stock, Dieter / Steffan, Birgit / Preuß, Rainer (1992): "Evaluierung des Sprachsynthesesystems HADIFIX." In *Fortschritte der Akustik, DAGA-92* (DEGA, Oldenburg)

Portele, Thomas / Stöber, Karl-Heinz / Meyer, Horst / Hess, Wolfgang (1996): "Generation of multiple synthesis inventories by a bootstrapping procedure." In *Proc. Intern. Conf. Spoken Language Processing* (ICSLP-96), Philadelphia, PA, USA

Rinscheid, Ansgar (1995): "Adaption von Sprache mit Hilfe von Merkmalskarten." In *Fortschritte der Akustik – DAGA 95* (DEGA, Oldenburg), 1031-1035

Rinscheid, Amsgar (1996): "Voice conversion base on topological feature maps andtime-variant filtering." In *Proc. Intern. Conf. Spoken Language Processing* (ICSLP-96), Philadelphia, PA, USA

Rodet, Xavier (1977): Analyse du signal vocal dans sa représentation amplitude-temps; synthèse de la parole par règles [Thèse d'Etat, Univ. Pierre et Marie Curie (Paris 6), F-91190 Gif-sur-Yvette]

Rodet, Xavier (1980): Time-domain formant-wave-function synthesis. In *Spoken language generation and understanding.* Proceedings of the NATO Advanced Study Institute held at Bonas, France, June 23 - July 5, 1979; ed. by J. C. Simon (Reidel, Dordrecht), 429-440

Rühl, Hans-Wilhelm (1989): "Sprachsynthese aus Text – Verfahren und Systeme für die deutsche Sprache." In *Fortschritte der Akustik, DAGA-89* (DEGA, Oldenburg), 105-120

Ruske, Günther / Schotola, Thomas (1978): "An approach to speech recognition using syllabic decision units."In *Proc. IEEE ICASSP-78* (IEEE, New York), 722-725

Schroeter, Juergen / Sondhi, Man Mohan (1992).: "Speech coding based on physiological models of speech production." In *Advances in speech signal processing;* ed. by M. M. Sondhi and S. Furui (Marcel Dekker, New York), 231-268

Sendlmeier, Walter F. (1991): "Wie testet man Hörverstehen? Eine kritische Analyse sprachaudiometrischer Testverfahren." In *Beiträge zur angewandten und experimentellen Phonetik;* hg. von W. Hess und W. F. Sendlmeier. Beihefte zur Z. für Dialektologie und Linguistik, Band 72 (Franz Steiner, Stuttgart), 83-101

Sendlmeier, Walter F. / Holzmann, Ulrike (1991): "Sprachgütebeurteilung mit Passagen fließender Rede." In *Fortschritte der Akustik, DAGA 91* (DEGA, Oldenburg)

Silverman, Kim / Beckman, Mary / Pitrelli, J. / Ostendorf, Mari / Wightman, C. / Price, Patti / Pierrehumbert, Janet / Hirschberg, Julia (1992): "ToBI: a standard for labelling English prosody." In *Proc. Intern. Conf. Spoken Language Processing* (ICSLP-92), Banff, Alberta, Canada

Sorin, Christel (1994): "Towards high-quality multilingual text-to-speech". In *Progress and prospects of speech research and technology*, ed. by H. Niemann, R. De Mori, and G. Hanrieder (infix, St. Augustin), 53-62

Sotscheck, Jochem (1982): "Ein Reimtest für Verständlichkeitsmessungen mit deutscher Sprache als ein verbessertes Verfahren zur Bestimmung der Sprachübertragungsgüte." Der Fernmeldeingenieur *36* (4/5), 1-45

Spiegel, Murray / Altom, Mary Jo / Macchi, Marian / Wallace, Karen (1989): A monosyllabic test corpus to evaluate the intelligibility of synthesized and natural speech. In *Proceedings of the ESCA workshop on Speech Input / Output Assessment and Speech Databases,* Noordwijkerhout, The Netherlands, September 1989 (Institute of Phonetic Sciences, Amsterdam)

Sproat, Richard / Olive, Joseph (1994/96): "A modular architecture for multi-lingual text-to-speech." In *Proc. of the Second ESCA/IEEE Workshop on Speech Synthesis*, New Paltz,NY, USA, 1994; to appear in *Progress in speech synthesis*, ed by J. Van Santen *et al.* (Springer, New York, Nov. 1996)

Stock, Dieter (1991): "P-TRA – eine Programmiersprache zur phonetischen Transkription." In *Beiträge zur angewandten und experimentellen Phonetik;* hg. von W. Hess und W. F. Sendlmeier. Beihefte zur Z. für Dialektologie und Linguistik, Band 72 (Franz Steiner, Stuttgart), 222-231

Taylor, Paul / Isard, Amy (1995): "SSML: a speech synthesis markup language." In *Proc. 2nd SPEAK!-Workshop* (GMD/IPSI, Darmstadt)

Terken, Jacques (1991): "Fundamental frequency and perceived prominence of accented syllables." J. Acoust. Soc. Am. *87*, 1768-1776

Traber, Christof (1992): "F0 generation with a database of natural F0 patterns and with a neural network." In *Talking machines: theories, models, and designs*, ed. by G. Bailly and C. Benoit (North-Holland, Amsterdam), 287-304

Traber, Christof (1993): "Syntactic processing and prosody control in the SVOX TTS system for German." In *Proc. EUROSPEECH-93* (catalyst consult, Berlin), 2099-2102

Traber, Christof (1995): *SVOX: The implementation of a text-to-speech system for German.* (Diss., ETH Zürich; TIK-Schriftenreihe Nr. 7; vdf Hochschulverlag, ETH Zürich)

Traber, Christof (1996): "Datengesteuerte Prosodiegenerierung mittels automatischer Lernverfahren." In *Fortschritte der Akustik, DAGA 96* (DEGA, Oldenburg), 86-89

Van Santen, Jan P. H. (1993): "Timing in text-to-speech systems." In *EUROSPEECH-93,* Berlin, Germany (catalyst consult, Berlin), 1397-1404

Van Santen, Jan P. H. (1994): "Assignment of segmental duration in text-to-speech synthesis." Computer Speech and Language *8*, 95-128

Van Santen, Jan P. H. / Sproat, Richard / Olive, Joseph / Hirschberg, Julia (eds.) (to appear in November 1996): *Progress in speech synthesis* (Springer, New York)

Wahlster, Wolfgang (1993): "VERBMOBIL – Translation of face-to-face dialogs." In *EUROSPEECH-93, Opening and Plenary Sessions,* Berlin, Germany (catalyst consult, Berlin), 29-38

Wesenick, M. Barbara / Schiel, Florian (1994): "Applying speech verification to a large data base of German to obtain a statistical survey about rules of pronunciation." In *Proc. Intern. Conf. Spoken Language Processing* (ICSLP-94), Yokohama, Japan, 279-282

Whalen, D. H. (1990): "Coarticulation is largely planned." J. Phonetics *18*, 3-35