Spracherkennung und Prosodie

E. Nöth, A. Batliner, A. Kießling, R. Kompe,
F. Gallwitz, V. Warnke, H. Niemann

Erst seitdem sich die automatische Sprachverarbeitung der Spontansprache und weniger restringierten Aufgabenstellungen zugewandt hat, ist der Einsatz der Prosodie wirklich sinnvoll geworden. Wir beschreiben im einzelnen die Gründe dafür und zeigen an der Integration der Prosodie in das automatische Übersetzungssystem Verbmobil, daß dieser Einsatz nicht nur sinnvoll, sondern auch erfolgreich ist.

1. Die Erkennung spontaner Sprache

Die Entwicklung der Forschung zur automatischen Spracherkennung im letzten Jahrzehnt ist geprägt durch eine fast ausschließliche Verwendung statistischer Ansätze (vorwiegend Hidden Markov Modelle, vgl. "Das aktuelle Schlagwort" in diesem Heft, aber auch Neuronale Netze) sowie durch eine verstärkte Zuwendung zu komplexeren Aufgaben: größerer Wortschatz, Sprecherunabhängigkeit, Spontansprache. Die kommerziell orientierten Systeme konzentrieren sich dabei natürlich noch auf relativ einfache Applikationen (sprecherabhängige Diktiersysteme, sprecherunabhängige Erkennung von sehr beschränkten Wortschätzen wie Zahlen oder Befehle). Die grundlagenorientierten Systeme wagen sich schon an sprecherunabhängige Übersetzung von Spontansprache (vgl. die Beiträge in diesem Heft zu Verbmobil).

Das Grundprinzip der statistischen Spracherkennungssysteme ist bei nahezu allen weltweit verwendeten Systemen gleich: Jedes Wort im Erkennungswortschatz wird durch ein geeignetes akustisches Modell, i.d.R. ein Hidden Markov Modell, repräsentiert, wobei sich die statistischen Parameter des Modells anhand von vorliegendem Sprachmaterial schätzen ("trainieren") lassen. Auf Grundlage dieses Modells läßt sich nun

Dr.-Ing. Elmar Nöth ist seit 1985 Mitarbeiter am Lehrstuhl für Mustererkennung der Universität Erlangen-Nürnberg (seit 1990 als Akademischer Rat). Seine Fachgebiete sind die Verarbeitung prosodischer Information, multilinguale Spracherkennung und semantische Verarbeitung mit stochastischen Mitteln.

Dr. phil. Anton Batliner ist seit 1997 wissenschaftlicher Mitarbeiter in der Forschungsgruppe des Lehrstuhls für Mustererkennung, wo er sich innerhalb der Sprachgruppe mit der Verwendung prosodischer Information in der automatischen Sprachverarbeitung beschaftigt.

Die anderen Autoren sind neben dem Lehrstuhlinhaber Professor Dr.-Ing. Heinrich Niemann weitere Mitarbeiter der Sprachgruppe des Lehrstuhls für Mustererkennung. Dr.-Ing Andreas Kießling ist heute bei Ericcson Eurolab in Nürnberg, Dr.-Ing Ralf Kompe am Sony Stuttgart Technology Center in Fellbach.

zu jedem Abschnitt eines Sprachsignals und zu jedem Wort die Wahrscheinlichkeit berechnen, daß an dieser Stelle tatsächlich das jeweilige Wort gesprochen wurde. Zudem lassen sich anhand von Textkorpora sogenannte Sprachmodelle schätzen, die die Wahrscheinlichkeit für das Auftreten bestimmter Wortfolgen berechnen. Das Problem der Spracherkennung wird nun so gelöst, daß mittels geeigneter Verfahren auf der Grundlage der akustischen Wortmodelle und des grammatischen Sprachmodells zu einem gegebenen Sprachsignal aus der Vielzahl aller möglichen Wortketten die wahrscheinlichste Wortsequenz bestimmt wird.

Die größere Komplexität der Spontansprache ist zum einen bedingt durch die nicht-kanonische (dialektale, verschliffene) Aussprache der Wörter, zum anderen durch die im Verhältnis zur Schriftsprache ‚freiere' - zumindest anders beschaffene - und auch oft ungrammatischere Syntax. Hinzu kommen Zögerungsphänomene (z.B, "äh", "hm"), nonverbale Geräusche wie Husten und Lachen sowie Wortabbrüche und Versprecher. Durch geeignete Modellierung solcher Phänomene im Worterkenner läßt sich die Zahl der hierdurch verursachten Erkennungsfehler deutlich reduzieren; dennoch stellt sich die Erkennung spontaner Sprache gegenüber der Erkennung gelesener Sprache als ein wesentlich schwierigeres Problem dar. Bild 1 zeigt die weltweite Entwicklung der Wortfehlerrate (Ordinate) in verschiedenen Systemen von 1988 bis 1996 (Abszisse). Die Wortfehlerrate gibt die Anzahl der vom Worterkenner im Mittel falsch erkannten, eingefügten oder ausgelassenen Wörter im Verhältnis zur Zahl der tatsächlich gesprochenen Wörter an; beispielsweise ist bei der Erkennung einer Äußerung von 10 Wörtern und einer Wortfehlerrate von 20% im Mittel mit 2 Fehlern zu rechnen. Die entscheidenden Faktoren für die Höhe der Fehlerrate sind dabei der Spontaneitätsgrad der Korpora (je spontaner, desto höher die anfängliche Fehlerrate), der Umfang (je größer das Korpus, desto besser die Möglichkeit der Modellierung mit einem Sprachmodell sowie die Möglichkeit der akustischen Modellierung), und die Laufzeit der Projekte (je länger, desto besser die Modellierung). Einzelheiten zu den Systemen finden sich z.B. in Kuhn (1995). Spontansprachliche Korpora sind ATIS *(Air Travel Information System), VERBMOBIL* und *SWITCHBOARD*. Die Fehlerrate von Verbmobil (2300 Wörter in der Domäne Terminabsprache, Mensch-Mensch-Kommunikation) konnte z.B. innerhalb von 3 Jahren von ca. 50% auf 12,5% reduziert werden. Sie liegt damit aber natürlich noch weit über der Fehlerrate von 0.8% bei den abgelesenen, kontinuierlich gesprochenen Zifferfolgen der TI-Digits-Stichprobe.

In all diesen Systemen gibt es eine Interaktion von akustischen Modellen mit Sprachmodellen, wobei jeweils die entprechende Einheit - der Laut bzw. das Wort - möglichst gut modelliert werden sollen. Wegen der artikulatorischen Abhängigkeiten der Laute von ihrer Umgebung und wegen der syntaktischen (syntagmatischen) Restriktionen des linken und rechten Kontextes eines Wortes im Satzzusammenhang werden die Einheiten nicht isoliert, sondern im Kontext modelliert. In der Akustik etwa als sog. Bi-, Tri- oder Polyphone, in der Sprachmodellierung als Bi-, Tri- oder Polygramme (jeweils zwei, drei oder mehr Einheiten werden zusammen betrachtet). Das Sprachmodell schränkt in der Spracherkennung die Freiheitsgrade des akustischen Modells ein. Dies soll kurz an Bild 2 erläutert werden. Dort ist ein Ausschnitt aus einem Worthypothesengraphen (WHG) dargestellt, so wie er von einem akustischen Klassifikator erzeugt wird. Die gesprochene Wortkette sei: ‚*... dachte noch in den nächsten ...*'; als andere Abfolgen, die ähnlich klingen und eventuell sogar akustisch besser bewertet sind, finden sich: ‚*... dachte noch ihnen nächsten ...*', ‚*... dachte noch in in den nächsten ...*', ‚*... dachte noch den der nächsten ...*' oder ‚*... dachte noch in in der nächsten ...*'. Diese Alternativen sind aber alle ungrammatisch, kommen

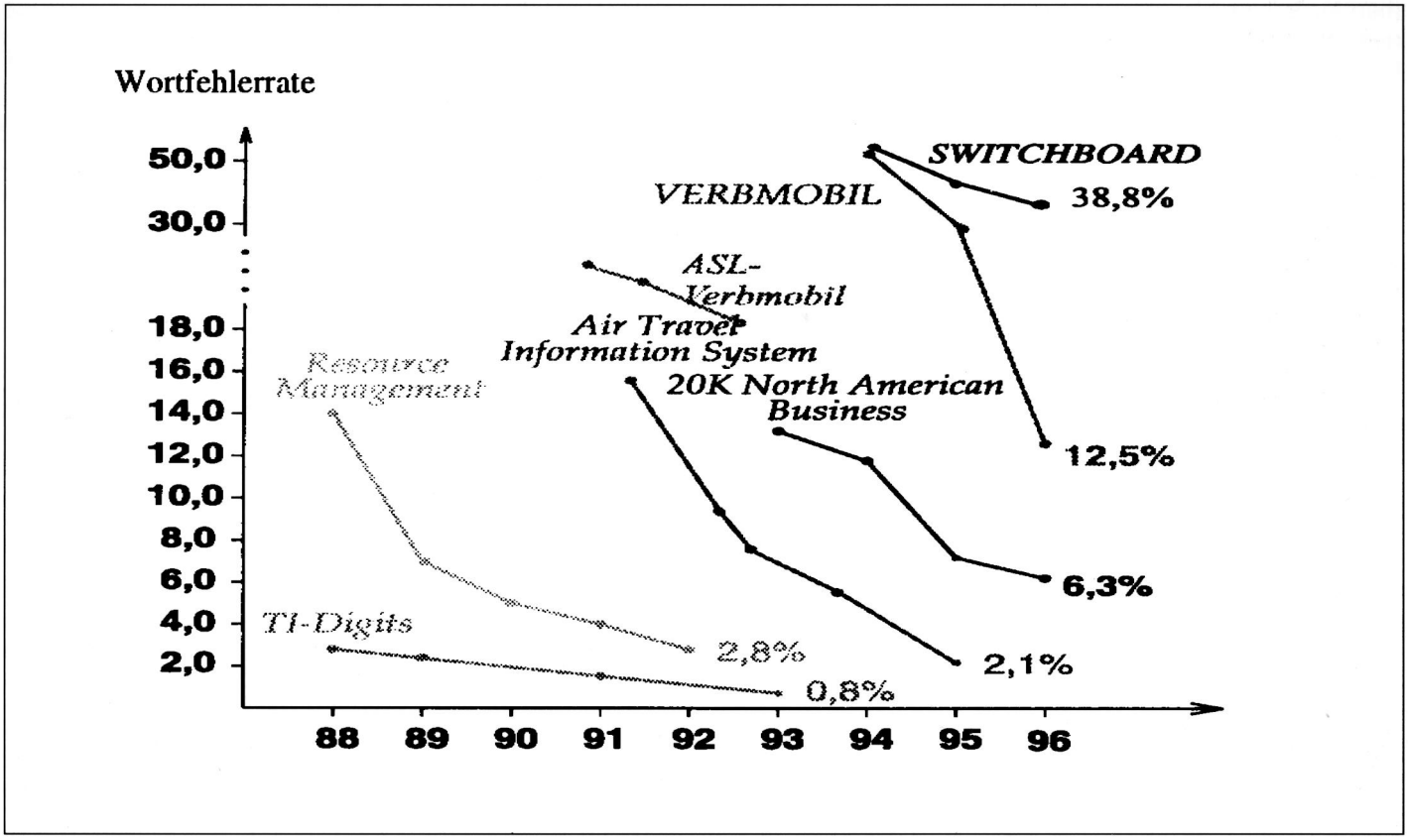

Bild 1: Entwicklung der Wortfehlerrate - weltweit

sehr selten (in spontaner Sprache) oder gar nicht (in geschriebener Sprache) vor und haben deshalb eine sehr niedrige a priori Wahrscheinlichkeit. Eine entsprechene Umgewichtung durch das Sprachmodell bewirkt, daß unwahrscheinliche Wortfolgen schlecht und wahrscheinliche besser bewertet werden. Damit nähern sich die besten Wortketten des WHG der gesprochenen, korrekten Wortfolge an. Die akustische und grammatische Modellierung in der statistischen Spracherkennung ist in Schukat-Talamazzini (1995) und in den dort zitierten Arbeiten genauer dargestellt.

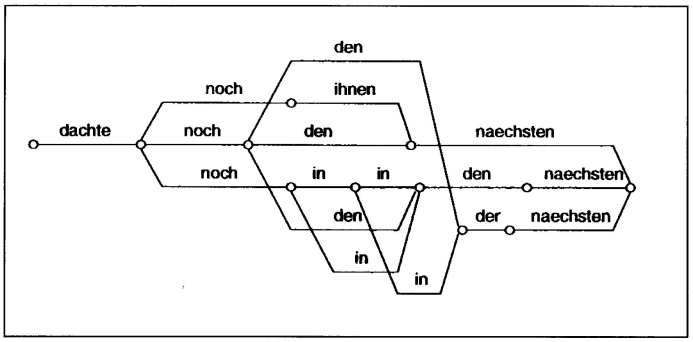

Bild 2: Ausschnitt aus einem Worthypothesengraphen

2. Prosodie und automatische Sprachverarbeitung

Die Prosodie beschäftigt sich mit suprasegmentalen (lautübergreifenden) sprachlichen Ereignissen. Diese Ereignisse überlagern sprachliche Einheiten, die mehr als einen Laut umfassen, also Silben, Wörter, Phrasen, Sätze, usw. Zur spektralen Dimension zählen Klangfarbe, Tonhöhe, Stimmlage und Stimmqualität, zur Intensität Lautheit, und die zeitliche Dimension umfaßt Pausensetzung, Dauerverhältnisse, Rhythmus, Sprechgeschwindigkeit und Tempo. Schon Lea (1980) und Vassiere (1988) haben die Verwendung prosodischer Information in der automatischen Sprachverarbeitung (ASV) gefordert. Das zentrale Problem im Zusammenhang mit der rechnergestützten Analyse prosodischer Information ist allerdings der hohe Komplexitätsgrad, der im wesentlichen durch folgende Faktoren verursacht wird: Beeinflussung durch segmentale Information, Interferenzen der unterschiedlichen prosodischen Funktionen, Interaktion der verschiedenen prosodischen Parameter, Fakultativität der prosodischen Mittel sowie sprecher- bzw. sprachenspezifische Faktoren. Die Komplexität ist aber nur einer der Gründe, warum Prosodie noch immer vergleichsweise selten in ASV-Systemen eingesetzt wird. Ein weiterer Grund ist, daß es bislang nur sehr wenige Anwendungen gibt, für die die bedeutungsunterscheidenden prosodischen Klassen auf der Hand liegen; für die übrigen Applikationen stellt bereits die Festlegung geeigneter Klassen ein Problem dar.

Im allgemeinen läßt sich die prosodische Analyse in folgende Schritte untergliedern: die Zerlegung einer Äußerung in prosodische Einheiten, die automatische Extraktion prosodischer Merkmale aus dem Sprachsignal über den entsprechenden Zeitintervallen und die Abbildung der Merkmalvektoren auf prosodische Klassen durch automatische Klassifikationsverfahren. Da zwischen der prosodischen und der zugrundeliegenden lautsprachlichen Information ein enger Zusammenhang besteht, liegt es nahe, die prosodischen Einheiten im Sprachsignal durch automatische Zeitzuordnung der gesprochenen Wortkette mit Hilfe eines Worterkenners zu lokalisieren. Eine genaue Darstellung dieser Schritte sowie der prosodischen Merkmale und ihrer Auswahl findet sich in Kießling (1997).

3. Prosodie und Disambiguierung

Auch wenn potentiell die Prosodie an vielen Stellen in einem System eingesetzt werden kann, so bietet es sich doch an, sie genau da einzusetzen, wo sie auch die wohl größte Rolle in der zwischenmenschlichen Kommunikation besitzt: bei der Segmentierung und Disambiguierung von Äußerungen. In den bisherigen Anwendungen waren allerdings die Äußerungen der Benutzer meist so kurz, daß diese Funktionen der Prosodie überhaupt nicht zum Tragen kamen. So ist zum Beispiel die durchschnittliche Länge eines Redebeitrags in einem Feldexperiment mit einem Zugauskunftssystem 3.5 Wörter, vgl. Eckert (1995). Verbmobil ist dagegen fast das einzige System, in dem ‚real life' Sprache in einer Mensch-Mensch-Kommunikation verarbeitet werden muß. 70% der Redebeiträge enhalten mehr als einen Satz; im Durchschnitt ist ein Turn 20 Wörter lang. Auch die in Spontansprache häufigen elliptischen Konstruktionen sowie Abbrüche und Neuansätze tragen zur Komplexität und Ambiguität bei. Mit deutlichem Erfolg kann Prosodie allerdings erst in der Verstehensphase, noch nicht in der Erkennungsphase, eingesetzt werden. Deshalb zeigt sich die Bedeutung der Prosodie erst in einem System wie Verbmobil, das eines der ersten Systeme ist, in dem eine end-to-end Evaluation das Optimierungskriterium darstellt, und in dem auch eine tiefe linguistische Analyse durchgeführt wird.

An einem einfach aussehenden Beispiel wollen wir nun die unterschiedlichen Funktionen der Prosodie bei der Disambiguierung von Redebeiträgen erläutern. Nehmen wir an, ein Sprecher sagt:

Können wir uns AUCH treffen? Vielleicht am nächsten Dienstag? Geht das?

Wörter, die im Hauptakzent stehen, sind durch Großschreibung gekennzeichnet. Wenn wir nur Worterkennung und Syntax zur Verfügung haben, so kann dieser Redebeitrag nicht von dem - wortidentischen - folgenden unterschieden werden:

Können wir uns auch TREFFEN vielleicht? Am nächsten Dienstag geht das.

In der folgenden Darstellung dieses Turns zeigen die senkrechten Striche potentielle Segmentierungsgrenzen an, die zusätzlich noch hinsichtlich des Satzmodus (Frage/Nicht-Frage) ambig sind:

Können wir uns auch treffen | vielleicht | am nächsten Dienstag | geht das |

Ein nachgestelltes Satzadverb wie *vielleicht* ist in der Schriftsprache zwar ungrammatisch, kommt aber in dieser Position in Spontansprache häufig vor. Je nach Tonverlauf am Ende eines Segmentes handelt es sich um eine Frage (steigender Verlauf) oder um eine Aussage (fallender Verlauf). Je nachdem, ob *wir* bzw. *auch* besonders betont sind oder nicht, oder ob die Akzentuierung auf *treffen* liegt, ändern sich Semantik und Präsupposition des Redebeitrags: Ein unbetontes *auch* indiziert als Alternative zu *treffen* etwa *telefonieren, schreiben,* ein betontes *auch* oder *wir* indiziert eine andere Alternative, nämlich die, daß sich andere Leute schon treffen, und wir das

auch tun sollten. Erschwert wird die Aufgabe der Spracherkennung natürlich dadurch, daß der Satz, fließend gesprochen, ja eher wie folgt aussieht, d.h., daß zuerst die Wortgrenzen bestimmt werden müssen:

KönnenwirunsauchtreffenvielleichtamnächstenDienstaggehtdas

4. Prosodie und Segmentierung

Ambige Fälle wie der im letzten Abschnitt diskutierte kommen immer wieder vor, sind allerdings nicht gerade der Regelfall. Dies gilt in noch stärkerem Maß für diffizilere Fälle, wie sie etwa die unterschiedliche Klammerung (PP-Attachment) und damit unterschiedliche Interpretation bei den beiden folgenden Sätzen mit identischer Wortfolge, aber anderer Bedeutung, darstellt:

Er sah (den Mann mit dem Fernglas) vs.
Er (sah den Mann) (mit dem Fernglas)

Im ersten Fall sah er einen Mann, der ein Fernglas trug, im zweiten Fall benutzte er das Fernglas, um den Mann zu sehen. Solche linguistisch interessanten Fälle sind in der Praxis marginal und werden u.E. eher durch Welt- oder Kontextwissen aufgelöst als durch eine unterschiedliche prosodische Indizierung. Weltwissen disambiguiert z.B. die Lesarten der beiden folgenden Sätze in die angegebene Klammerung:

Er schlug (den Mann mit dem Strohhut) vs.
Er schlug (den Mann) (mit dem Knüppel)

Bei weitem wichtiger ist die Aufgabe der Prosodie, der gesprochenen Sprache eine allgemeine Struktur zu geben. Sie ist das Gliederungsmittel, das für die geschriebene Sprache die Interpunktion sowie das weitere Layout eines Textes (Absätze, unterschiedliche Schriftgröße, Hervorhebungen durch Fettdruck etc.) darstellt. Allerdings ist diese Information grundsätzlich nicht kategorial, im Sinne eines Entweder-Oder zu verwenden, sondern als - mehr oder weniger hohe - Wahrscheinlichkeit für die Richtigkeit einer Zerlegung. Bild 3 verdeutlicht diese Funktion der Prosodie durch eine Übertragung auf die geschriebene Sprache. Links sieht man eine ungegliederte Abfolge von Wörtern, also den **reinen Text**; in der Mitte sieht man das "Gegenteil", nämlich nichts anderes als die **reine Struktur**, da die Buchstaben durch Kästchen ersetzt sind. Rechts steht die gewohnte Darstellung, nämlich ein **strukturierter Text**. In Analogie kann man sich eine **reine Wortfolge** vorstellen, so wie sie völlig monoton etwa von einer veralteten Synthese produziert wird (links). Eine **"reine Prosodie"** kann z.B. instrumentell dadurch erzeugt werden, daß man mit geeigneten Verfahren die Lautinformation wegfiltert (Mitte). Der strukturierte Text entspricht der natürlich gesprochenen, **prosodisch strukturierten** Sprache.

Man beachte, daß der reine Text ja schon aufbereitet ist, da sich zwischen den Wörtern Leerzeichen befinden. Trotzdem liest er sich sehr schlecht, da vom Leser die zusätzliche Leistung verlangt wird, während des Lesens die Textstruktur zu ermitteln. Die reine Struktur sagt natürlich nichts über den Inhalt aus; sie kann aber einem Leser, der diese Art von Text gewohnt ist, der also weiß, wie wissenschaftliche Artikel aufgebaut sind, sehr viele Hinweise darauf geben, wo welche Textteile zu finden sind: die Überschrift, die Autoren, die Fußnote links unten, usw.

Es wäre allerdings irreführend, wenn man den Beitrag der Prosodie zur Strukturierung in der ASV völlig analog zu ihrem Beitrag beim Verstehen des Gehörten durch den Menschen sehen würde; der Beitrag ist noch größer. Dem Menschen steht normalerweise die gesprochene Wortkette zur Verfügung, und wenn er nicht alle Wörter versteht, so handelt es sich doch fast immer nur um wenige alternative Möglichkeiten der Interpretation. In der automatischen Spracherkennung ist es nicht der Regelfall, sondern die Ausnahme, wenn die gesprochene Wortkette mit der besten Wortkette des WHG identisch ist. Der Mensch kann also schon während des Verstehensprozesses falsche Alternativen ausschalten. In der automatischen Spracherkennung sind dagegen z.B. 12 alternative Hypothesen pro Wort eine gängige Größe. Die Zahl der falschen

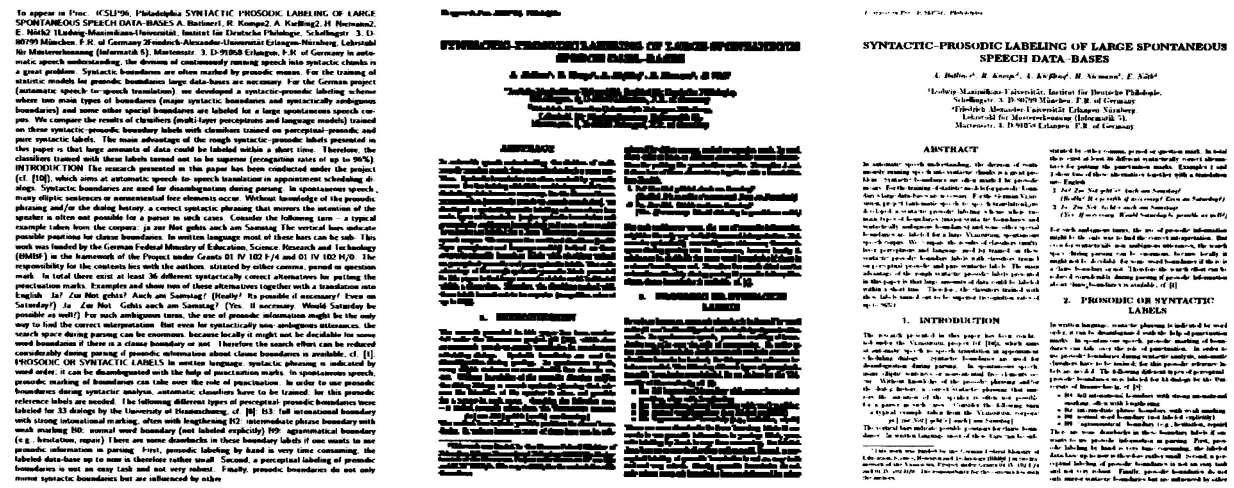

Bild 3: reiner Text (links), reine Struktur (Mitte), strukturierter Text (rechts)

Hypothesen ist also um eine Größenordnung höher als die der richtigen Hypothesen. Das Suchproblem ist deshalb enorm. Aus diesem Grund wird prosodische Information in Verbmobil im Augenblick hauptsächlich bei der syntaktischen Analyse für die Suche durch den Wortgraphen nach dem besten Parse (der besten syntaktischen Strukturanalyse) eingesetzt. Dabei werden keine harten Entscheidungen getroffen. Partielle Parses werden in einer Agenda nach einer Gewichtung angeordnet, die die prosodische Wahrscheinlichkeit für eine syntaktische (Teil-)Satzgrenze angibt. In jedem Schritt der Suche wird der jeweils beste partielle Parse verwendet. Prosodische Information beschleunigt also die Suche nach dem besten kompletten Parse. Da in realistischen Systemen die Suche nicht zu lange dauern darf und deshalb nach einer gewissen Zeit abgebrochen werden muß, erhöht diese Beschleunigung gleichzeitig die Erkennungsrate des Syntaxmoduls.

Experimente in Verbmobil haben gezeigt, daß mit der Kombination eines Neuronalen Netzes (Multi-Layer-Perceptron) mit einem Sprachmodell, das die möglichen Abfolgen von Wörtern und syntaktischen Grenzen modelliert, syntaktische Grenzen und Nicht-Grenzen mit einer Erkennungsrate von bis zu 94% richtig unterschieden werden. Ohne den Einsatz der Prosodie müssen 138 Alternativen untersucht werden, mit Prosodie sind es 6. Die Parsezeit ohne Prosodie beträgt im Mittel 39 Sekunden, mit Prosodie 3 Sekunden. Dies bedeutet eine Reduktion der Parse-Alternativen um durchschnittlich 96% und der Parse-Zeit um durchschnittlich 92%. Damit zeigt sich, daß erst der Einsatz der Prosodie Verbmobil zu einem System mit für einen Benutzer akzeptablen Verarbeitungszeiten macht; zu Einzelheiten vgl. Kompe (1997). Die syntaktische Verarbeitung in Verbmobil ist in Block (1997) dargestellt, ein allgemeiner Überblick über Verbmobil findet sich in Bub (1997).

5. Segmentierung und Übersetzung

Beim Übersetzen von gesprochener Sprache können mindestens vier unterschiedliche Ansätze gewählt werden, wobei auch hybride Ansätze, welche auf diesen grundlegenden Ansätzen basieren, möglich sind (siehe hierzu auch den Beitrag von Menzel und Quantz in diesem Heft):

Flache Übersetzung: Die Wortsequenz wird in eine Sequenz von Kommunikationsschritten zerlegt. Die Bedeutung der Kommunikationsschritte wird in der Zielsprache dargeboten, unabhängig von der exakten Wortfolge, mit der die Kommunikationsschritte in der Quellsprache realisiert wurden.

Tiefe Übersetzung: Es wird eine linguistische Analyse der Wortsequenz durchgeführt. Die linguistischen Strukturen werden in die Zielsprache übergeführt.

Stochastischer Ansatz: Anhand eines zweisprachigen Korpus werden statistische Modelle trainiert, welche für eine Wortsequenz in der Quellsprache eine Wortsequenz in der Zielsprache vorhersagen.

Übersetzungsgedächtnis: Alle Sätze oder Teilsätze, welche das System übersetzen kann, sind in einer Liste gespeichert.

Allen Ansätzen ist gemeinsam, daß sie eine Zerlegung des Sprachsignals in kleinere Einheiten erfordern, wobei der Detailliertheitsgrad der Zerlegung vom gewählten Ansatz abhängt. Dies läßt sich anhand der tiefen und der flachen Übersetzung, so wie sie auch im Übersetzungssystem Verbmobil durchgeführt werden, gut erläutern: bei der flachen Übersetzung werden Kommunikationsschritte gesucht, die als Dialogakte bezeichnet werden. Ein solcher Akt kann z.B. eine Ablehnung oder eine Begründung einer Entscheidung sein, d.h. die Grenzen für die Zerlegung sind über die kommunikative Rolle der Einheiten gegeben; vgl. dazu Reithinger (1995). Bei der tiefen Analyse sind hingegen syntaktische Strukturen Ausgangspunkt der Zerlegung, also etwa Grenzen zwischen Haupt- und Nebensatz oder zwischen zwei Sätzen.

Im Verbmobil-Prototypen werden tiefe und flache Übersetzung parallel durchgeführt; wenn die tiefe Übersetzung zu einem Ergebnis kommt, wird dieses genommen, wenn nicht, wird auf das Ergebnis der flachen Analyse zurückgegriffen. Das Grundproblem ist für beide Arten der Analyse das gleiche: Die Anzahl der möglichen Alternativen ist sehr groß und die Zeit, die für die Berechnung zur Verfügung steht, kurz. Ohne Berücksichtigung der prosodischen Information wäre die Wahl der vom Sprecher intendierten Alternative nur mit genügend Kontextinformation und Dialog(-schritt)-Wissen machbar. Dieses Wissen steht nicht zur Verfügung. In Verbmobil werden die für die flache Übersetzung notwendigen Dialogaktgrenzen von einem prosodischen Klassifikator zu 90% erkannt.

6. Ausblick

In der zweiten Phase von Verbmobil (1997-2000) sollen zumindest die folgenden Punkte zusätzlich bearbeitet werden:

(1) Weiterführende Integration der Sprachmodelle in die Suche des Parsers.

(2) Regelbasierte Bestimmung der Akzentposition in der Wortkette pro Akzentphrase, ausgehend von syntaktisch-prosodischen Grenzen und annotierten Wortarten (parts-of-speech).

(3) Übertragung auf andere Sprachen (Englisch, Japanisch)

Möglicherweise ist der Beitrag der Prosodie zur ASV bei der Segmentierung der gesprochenen Wortkette in verarbeitbare Einheiten – seien das nun Dialogakte, Translationseinheiten oder "klassische" syntaktische Phrasierungseinheiten - wirklich am größten. Allerdings stehen wir erst am Anfang der Integration der Prosodie in die ASV. Eine weitere Verbesserung der Erkenner- und Verstehensleistung durch ihren Einsatz ist daher sehr wahrscheinlich.

Danksagung

Diese Arbeit wurde vom Bundesministerium für Bildung, Wissenschaft, Forschung, und Technologie (BMBF) im Rahmen des Verbmobil-Projekts unter der Fördernummer 01 IV 701 K5 gefördert. Die Verantwortung für den Inhalt liegt bei den Autoren. Wir danken Thomas Kuhn und Peter Regel (Daimler Benz) für die Überlassung von Bild 1.

Literatur

Block (1997): Block, H. U.: The Language Component in Verbmobil. In: *Proc. ICASSP 1997,* München 1997: 79-82.

Bub (1997): Bub, T., Wahlster, W., Waibel, A.: Verbmobil: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation. In: *Proc. ICASSP 1997,* München 1997: 71-74.

Eckert (1995): Eckert, W., Nöth, E., Niemann, H., Schukat-Talamazzini, E.: Real Users Behave Weird - Experiences made collecting large Human-Machine-Dialog Corpora. In: P. Dalsgaard, L.B. Larsen, L. Boves, I. Thomsen (Hrsg.): *Proc. of the ESCA Tutorial and Research Workshop on Spoken Dialogue Systems*, Vigsoe 1995: 193-196.

Kießling (1997): Kießling, A.: *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung.* Aachen 1997.

Kompe (1997): Kompe, R., Kießling, A., Niemann, H., Nöth, E., Batliner, A., Schachtl, S., Ruland, T., Block, H.U.: *Improving Parsing of Spontaneous Speech with the Help of Prosodic Boundaries*. In: *Proc. ICASSP 1997,* München 1997: 811-815.

Kuhn (1995): Kuhn, T.: *Die Erkennungsphase in einem Dialogsystem.* Sankt Augustin 1995.

Lea (1980): Prosodic Aids to Speech Recognition. In: W.E. Lea (Hrsg.): *Trends in Speech Recognition,* Englewood Cliffs, New Jersey, 1980: 166-205.

Reithinger (1995): Reithinger, N., Maier, E., Alexandersson, J.: Treatment of Incomplete Dialogues in a Speech-to-speech Translation System. In: P. Dalsgaard, L.B. Larsen, L. Boves, I. Thomsen (Hrsg.): *Proc. of the ESCA Tutorial and Research Workshop on Spoken Dialogue Systems*, Vigsoe 1995: 33-36.

Schukat-Talamazzini (1995): *Statistische Spracherkennung.* In: KI, 1995: 7-9.

Vaissière (1988): Vaissière, J.: The Use of Prosodic Parameters in Automatic Speech Recognition. In: Niemann, H., Lang, M., Sagerer, G. (Hrsg.): *Recent Advances in Speech Understanding and Dialog Systems*, Berlin 1988: 71-99.