ÜBER DIE ANWENDBARKEIT EINFACHER PROSODISCHER REGELN FÜR DIE SPRACHSYNTHESE

R. BAKIS, E. H. ROTHAUSER, D. MAIWALD*

Strategien für die Erzeugung synthetischer Sprache müssen sich mit drei Problemen auseinandersetzen: a) Wahl der Elemente (Lautelemente oder -gruppen, Wörter, Wortgruppen), b) Verkettung von und Interpolation zwischen den Elementen, c) Berücksichtigung der prosodischen Suprastruktur. Zahlreiche Arbeiten haben bereits vielversprechende Teillösungen gebracht, die aber nur die beiden ersten Problemgebiete berücksichtigen und deshalb prinzipiell nur zu Sprache mit begrenzter Qualität führen können. Grundlegende Verbesserungen der Qualität erfordern eine quantitative Erfassung der prosodischen Strukturen, die in der Literatur noch nicht mit der für uns benötigten Genauigkeit beschrieben sind. In qualitativer Hinsicht wissen wir, daß die prosodischen Merkmale der Sprache vor allem durch den Tonhöhenverlauf und die zeitliche Struktur der Elemente gegeben sind. Wir suchen als Arbeitsmodelle einfache Algorithmen zur Bestimmung des zeitlichen Verlaufs der Parameter Tonhöhe und Dauer, um einen vorgegebenen Sprechausdruck zu erhalten. Durch spätere schrittweise Verfeinerung dieser Modelle, vielleicht unter Verwendung zusätzlicher Parameter, sollte ihre Verknüpfung zunächst mit der grammatischen Struktur eines Satzes möglich werden und auf einer anderen, vielleicht höheren Stufe, mit seinem semantischen Gehalt oder mit dem Emotionszustand des Sprechers.

Bei den genannten Untersuchungen verwenden wir für die Sprachsynthese einen Computer-gesteuerten Vocoder.[1] Dadurch haben wir die Möglichkeit, bei einem gegebenen Wort oder Satz die Tonhöhe ohne Beeinflussung der Dauer zu verändern und umgekehrt. In der primitivsten Stufe einer Synthese genügt es demnach, die prosodischen Merkmale Tonhöhe und Dauer für das jeweilige Sprachelement direkt vorzugeben. Hierzu ein einfaches Beispiel, bei dem als Sprachelement das Wort „ai" verwendet wurde. Bei wirklicher Sprache dürfen die Übergänge zwischen verschiedenen Tonhöhen jedoch nicht so abrupt gemacht werden, außerdem ist es bei der Sprachsynthese nicht sinnvoll, die Vielfalt aller möglichen Tonhöhen zur Verfügung zu stellen.

* IBM Forschungslaboratorium, 8803 Rüschlikon, Schweiz.

[1] E. H. Rothauser: The integrated vocoder and its application in computer systems. *IBM Journal* 10 (1966), 455—461.

Der Verlauf der Tonhöhe kann nach Öhman und Lindqvist[2]) bei normaler Sprache durch die Kombination dreier Komponenten angegeben werden, nämlich durch die Satzmelodie, eine Folge von Einzelwortmelodien und eine physiologisch bedingte Welligkeit (interaction ripple). Der Verlauf der Satzmelodie kann durch eine zeitliche Folge von nur wenigen, geglätteten Stufenfunktionen angenähert werden. Die Melodieempfindung wird nach Hockett[3]) durch die Verwendung von vier Tonhöhenniveaus und den drei Endungen „eben", „fallend" und „steigend" hinreichend gut nachgebildet. Um annehmbar klingende Sprache zu erzeugen, muß man auch ihre zeitliche Struktur beeinflussen. Einige Elemente müssen gedehnt, andere aber in ihrer Zeitdauer verkürzt werden. Wir haben gefunden, daß auch bei sehr unterschiedlich dauernden Elementen keine Glättung der Geschwindigkeitssprünge erforderlich ist. Nach Hockett[3]) ist die englische Sprache dadurch charakterisiert, daß bei gegebener Sprechgeschwindigkeit es ebenso lang dauert, um von einer betonten Silbe zur nächsten zu gelangen, ob keine oder viele Silben dazwischenliegen. Es wurde ein Programm entworfen, mit dem im Computer gespeicherte Einzelwörter zu Sätzen zusammengestellt und mit prosodischen Merkmalen versehen werden können, die wir in Anlehnung an die erwähnten Ergebnisse obiger Autoren formuliert haben. Durch die Verwendung ganzer Wörter als Sprachelemente können wir die Einzelwortmelodie und die oben beschriebene Welligkeit als gegeben voraussetzen. Fig. 1 zeigt am Beispiel des einfachen Satzes „Good morning class" die von uns verwendeten Notierungen und Algorithmen. Die oberste Zeile gibt den gewählten Satz wieder, wie er dem Rechner zugeführt wird, mit den Zahlen für das gewünschte Tonhöhenniveau in Halboktavschritten und den Symbolen für die Endungen. Zu

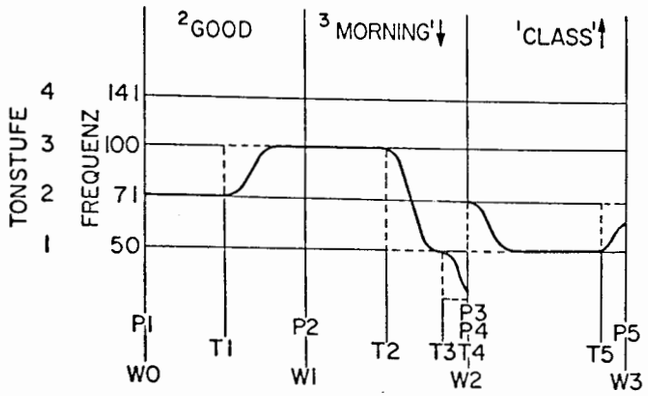

Fig. 1.

[2] S. Öhman und J. Lindqvist: Analysis-by-synthesis of prosodic pitch contours. *STL-QPSR*-4 1965 Royal Institute of Technology, Stockholm, 1—6.

[3] C. F. Hockett: A course in modern linguistics. The Macmillan Company, New York 1958.

Beginn eines Satzes und auch am Anfang jedes Makrosegmentes existiert ein bestimmtes Tonhöhenniveau, für das wir Nummer 2 gewählt haben.

Der Übergang von einem Niveau zum anderen wird geglättet und erfolgt jeweils in der Mitte zwischen den Punkten, bei denen eine Tonhöhenzahl angegeben ist (T_1, T_2 in Fig. 1). Die Übergänge T_3 und T_5 rühren von den Angaben über die gewünschte Endung her. Die Stufe beginnt immer 100 msec vor dem Zeitpunkt, für den die Endung angeschrieben ist. Der Algorithmus für Betonungen lautet: Die Sprechdauer zwischen zwei aufeinanderfolgenden Betonungszeichen wird auf einen konstanten Wert 0,8 sec transformiert. Zuvor wird die Zeitdauer für den unbetonten Teil um den Faktor 2 verkürzt.

Das unmittelbare Ergebnis der vorliegenden Arbeit ist die Erkenntnis, daß einfache prosodische Modelle zwar für eine erste, phänomenologische Beschreibung der Sprache ausreichen mögen, daß aber für eine Synthese von Sprache höchster Qualität wesentlich kompliziertere Modelle erforderlich sind. Wir hoffen, außerdem gezeigt zu haben, daß Arbeiten auf dem Gebiet der Sprachsynthese in ein Grenzgebiet fallen, das für Phonetik, Linguistik, Nachrichtentechnik und Datenverarbeitung gleichermaßen von Interesse ist.

DISCUSSION

Paulus:

Für die praktische Anwendung ist die vorgeschlagene Methode wegen ihrer Allgemeinheit unvorteilhaft. Die Erstellung der Daten erfordert vom Benützer gründliche Kenntnisse über prosodische Merkmale. Besteht die Absicht, die Benützung durch Verwendung von geeigneten „Prosodemen" (deren Anzahl überschaubar gering sein muß) zu erleichtern?

Maiwald:

ad Paulus:

Zum ersten Teil der Frage:

Um die Vielfalt der erzeugbaren prosodischen Muster für den Anwender einzuschränken und damit die Programmierarbeit für ihn zu vereinfachen, verwenden wir eine übergeordnete Notierung, wobei — vergleichbar dem Zusammenhang zwischen FORTRAN und Maschinensprache — durch einfache Symbole die nötigen Unterprogramme aufgerufen und durchgeführt werden.

Zum zweiten Teil:

Im Augenblick sind wir dabei, die grammatische Struktur der Sätze einzubeziehen, derart, daß für gleiche grammatische Strukturen gleiche prosodische Muster verwendet werden, unabhängig von den im Satz auftretenden Wörtern. Es ist leicht einzusehen, daß dieses Verfahren zu einer radikalen Einschränkung der Programmierarbeit für den Anwender führen kann. Die Ergebnisse sind recht ermutigend.