Messungen zur intra- und interindividuellen Vokalvarianz und ihre Repräsentation in Lautheitsmustern

F. KUNERT
(Institut für Elektroakustik, Technische Universität München)

Die gehörgerechte Vorverarbeitung des Sprachsignals bei der Automatischen Spracherkennung hat sich seit langem bewährt [1, 2]. Die spezifischen Lautheiten, die diese Vorverarbeitung nach Zwicker liefert, beinhalten wesentliche Informationen, die das menschliche Gehör aufnimmt. Da die bisherigen Spracherkennungsssysteme, die auf einer Vorverarbeitung nach dem Lautheitsmodell beruhen, fast ausnahmslos sprecherabhängig sind, soll hier die Auswirkung verschiedener Sprecher auf die Lautheitsmuster der Vokale untersucht werden.

Der untersuchte Wortschatz besteht aus 74 Wörtern, davon 68 verschiedene, aus denen elf sinnvolle Sätze gebildet wurden. Die Sätze beinhalten bei korrekter Aussprache 121 Vokale, davon 20 verschiedene einschließlich der Diphtonge. Das Sprachmaterial hat eine für die deutsche Sprache typische Vokalverteilung, wenn eine Etikettierung nach Duden zugrundegelegt wird. Die 11 Sätze wurden von jeweils 5 Männern und 5 Frauen im reflexionsarmen Raum aufgesprochen. Die Versuchspersonen wurden vor der Aufnahme angewiesen, in der für sie typischen Art zu sprechen, also weder besonders langsam noch übertrieben deutlich.

Das auf Band aufgezeichnete Sprachmaterial wurde in der in Abb. 1 skizzierten Anordnung weiterverarbeitet.

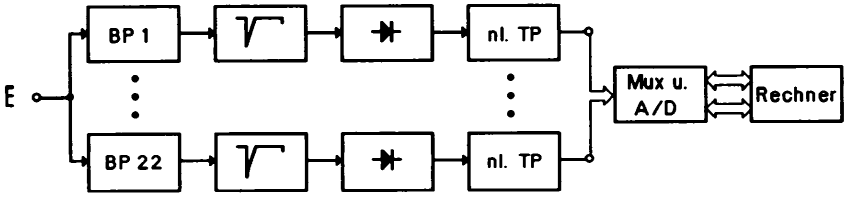

Abb. 1: *Blockschaltbild des Meßaufbaus.*

Die ersten vier Blöcke in Abb. 1 stellen die wesentlichen Stufen eines Lautheitsmodells nach Zwicker dar. Gezeichnet ist nur der erste und letzte Kanal, da alle Kanäle gleich aufgebaut sind. Sie unterscheiden sich nur in den Bandgrenzen der Bandpässe. In der ersten Stufe wird das Eingangssignal in 22 frequenzgruppenbreite Kanäle aufgeteilt. Die 22 Bandpässe erfassen einen Frequenzbereich von 50 Hz bis 12 kHz. Die den Filtern folgenden Radizierer transformieren die 22 Signale in die spezifischen Lautheiten N'. Nach der Gleichrichtung erfolgt in den speziellen "nichtlinearen Tiefpässen" eine zeitliche Bewertung der spezifischen Lautheiten, die der Nachverdeckung entspricht.

Die so gewonnenen spezifischen Lautheiten werden anschließend über einen Multiplexer zum A/D-Wandler geführt, mit 12 bit quantisiert und auf die Festplatte einer Workstation (HP 9000, Betriebssystem UNIX) geschrieben (ein Datensatz à 22 Werte pro ms).

Die eingelesenen Daten wurden graphisch aufbereitet, die Vokale lokalisiert und etikettiert.

Zur Lokalisierung der Vokale im Sprachmaterial ist die Summenlautheit, die man durch die Summation der spezifischen Lautheiten unter Berücksichtigung von spektraler Verdeckung und nichtlinearer Auffächerung der oberen Flanke erhält, gut geeignet. Zur Vermeidung von Fehlanzeigen aufgrund lauter Frikative wie /s/ oder lauter Nasale wurde die "modifizierte Lautheit" N_m [3] verwendet, bei der die unteren drei Kanäle sowie die oberen Kanäle ab Kanal 16 bei der Summation nicht berücksichtigt werden.

Durch das Weglassen der tiefen Kanäle werden die Nasale geschwächt, während das Fehlen der Frequenzen oberhalb von 3,5 kHz die Frikative unterdrückt. Glättet man die modifizierte Lautheit anschließend durch Tiefpaßfilterung, so verschwinden auch die kurzen Spitzen der Plosive.

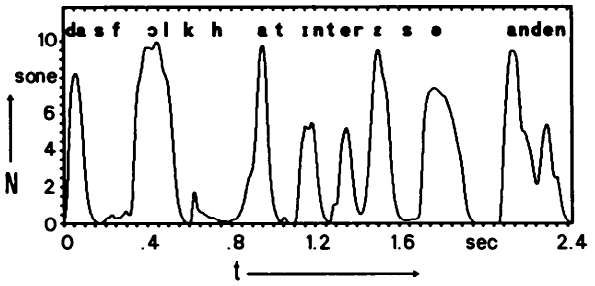

Abb. 2: *Summenlautheit mit Lautschrift (Sprecherin F1).*

In Abb. 2 ist für eine Sprecherin der Verlauf der geglätteten modifizierten Lautheit für einen Ausschnitt aus einem der Testsätze wiedergegeben. Im oberen Teil des Bildes ist die Lautschrift, wie sie sich durch eine Etikettierung nach Duden ergibt, eingezeichnet. Deutlich erkennbar sind die Vokale (durch die Maxima im Lautheitsverlauf) und die fast vollständige Unterdrückung der Frikative. Die kurzen Spitzen der Plosive sind durch die Glättung fast eingeebnet. Als problematisch erweisen sich lediglich der vokalische Konsonant /l/ und die Nasale, die trotz des Weglassens der drei unteren Kanäle noch zu einer auf Vokale hinweisenden Anzeige führen können. Dies kann in etwas "nachlässig" gesprochenen Passagen zu einer Vokalanzeige führen. Es zeigt sich, daß die geglättete modifizierte Lautheit bei allen zehn Sprechern gut zur Lokalisierung von Vokalen geeignet ist. Deshalb kann dieses Verfahren als sprecherunabhängig bezeichnet werden.

Bei der Etikettierung der so lokalisierten Vokale traten, unabhängig von Geschlecht und Herkunft der Sprecher, einige Abweichungen zwischen der Lautschrift nach Duden und den aufgezeichneten Testsätzen auf. So wurde der Murmellaut /ə/ vor Nasalen oder einem /l/ am Wortende häufig ausgelassen, wodurch die Nasale oder das /l/ am Wortende als Vokal angezeigt werden können. Auch wird das /ə/ vor dem Vibrant /r/ oft als offenes `a` unter Auslassung des nachfolgenden /r/ gesprochen.

Im Gegensatz zur deutschen Sprachstatistik ist in den untersuchten Sätzen nicht das `e` sondern das `i` der am häufigsten auftretende Vokal, der im Folgenden, stellvertretend für alle anderen Vokale, noch näher betrachtet wird.

Abb. 3 zeigt den Verlauf der spezifischen Lautheiten in den einzelnen Kanälen für ein typisches /ɪ/ über der Zeit (männlicher Sprecher). Es handelt sich um das `i` aus dem Wort `Verhältnissen`. Man erkennt noch einen Teil des vorhergehenden /n/ und den Anfang des nachfolgenden /s/.

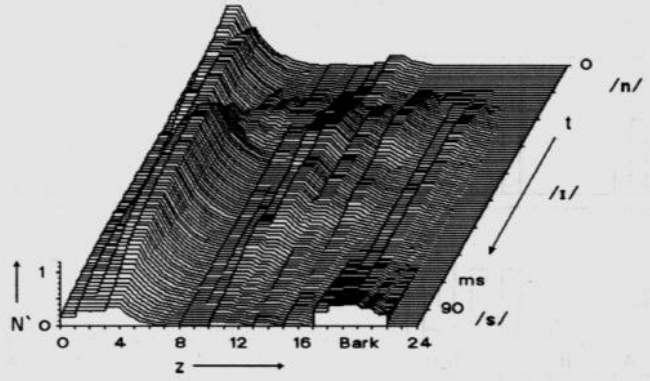

Abb. 3: *Lautheitsmuster des Vokals /ɪ/ im Wort "Verhältnissen" (Sprecher M1).*

Für alle `i` jedes Sprechers wurde durch einen Schnitt entlang der Tonheitsachse zum Zeitpunkt des Maximums der modifizierten Lautheit ein Muster des Vokals in der Lautheits–Tonheitsebene gewonnen. Die Wahl des Zeitpunktes ist unkritisch, da sich die Charakteristika eines Vokals über einen längeren Zeitraum kaum ändern.

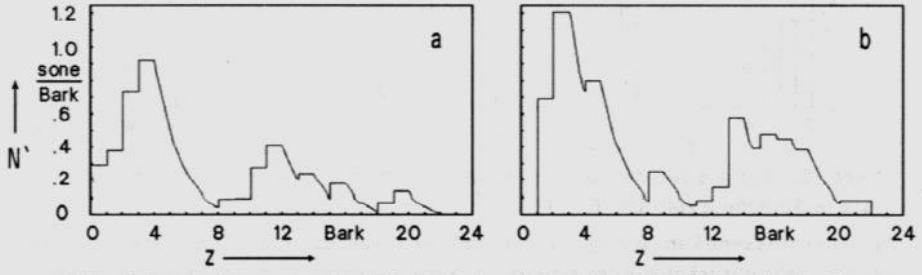

Abb. 4: Lautheitsmuster des Vokals /ɪ/. a) Sprecher M1, b) Sprecherin F1.

In Bild 4 sind exemplarisch 2 Muster abgebildet. Aus solchen Mustern wurden für jeden Sprecher die Häufigkeiten h der lokalen Maxima in den einzelnen Kanälen gebildet. Die Ergebnisse der so gewonnenen Daten sind für jeweils 4 Männer (M1-M4) und 4 Frauen (F1-F4) in Abb. 5 dargestellt.

In dieser Darstellung lassen sich Gemeinsamkeiten und sprechertypische Charakteristika gut erkennen. Bei allen Männern fällt der erste Formant in die Kanäle 3 oder 4, seltener in Kanal 5, während er bei den weiblichen Sprechern in den Kanälen 4, 5 und 6 liegt. Eine Ausnahme stellt Sprecherin F4 dar, deren erster Formant in den Kanälen 3-5 liegt. Bezüglich des zweiten Formanten kann man die männlichen Sprecher in 2 Gruppen einteilen, M1 mit M4 und M2 mit M3. Auch haben nur M1 und M4 häufig ein Maximum im Kanal 20 oder

21, was auch bei Sprecherin F2, allerdings in Kanal 21 und 22 der Fall ist. Wie schon in Abb. 4 erkennbar ist, sind die Maxima in Kanal 8 und 9 charakteristisch für das `i` von Sprecherin F1.

$$h = \frac{\Sigma \text{ lok. Max. im Kanal } i}{\text{Anzahl der Muster}}$$

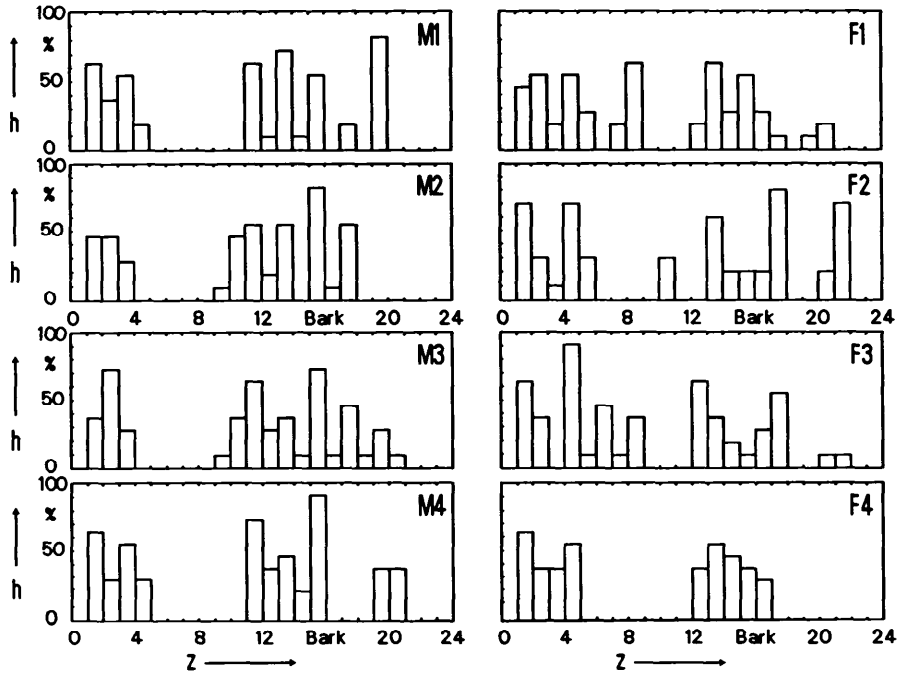

Abb. 5: *Häufigkeiten der lokalen Maxima in den einzelnen Kanälen für das /ɪ/ von 8 Sprechern (M1-M4, F1-F4).*

Zusammenfassend kann gesagt werden, daß die Vokale durch die geglättete modifizierte Lautheit sehr gut lokalisierbar sind und das Verfahren sehr robust in Bezug auf unterschiedliche Sprecher ist. Auch bilden sich sprechertypische Charakteristika gut in den Lautheitsmustern ab und können so vorteilhaft für eine Sprecheradaption eingesetzt werden.

Die Untersuchungen wurden im Rahmen des SFB 204 "Gehör", München, von der Deutschen Forschungsgemeinschaft unterstützt.

Literatur

[1] Zwicker, E., Terhardt, E., and Paulus, E. (1979) Automatic speech recognition using psychoakustic models. J. Acoust. Soc. Am. 65, 487-498.

[2] Ruske, G. (1988) Automatische Spracherkennung. Oldenbourg, München.

[3] Schotola, T. (1980) Silbenanlautende und Silbenauslautende Konsonantenfolgen als Entscheidungseinheiten für die automatische Spracherkennung. Diss. Techn. Universität München.