EINIGE AKUSTISCHE PARAMETER ZUR SPRECHERKLASSIFIKATION

MICHAEL KIRSTEIN

Schon seit längerer Zeit wird in der Akustischen Phonetik der Versuch unternommen, von den am Sprachsignal kommunikationstheoretisch zu unterscheidenden Klassen der inhalt-, situation- und sprechergebundenen Merkmale die letztere experimentell in den Griff zu bekommen und von den anderen zu isolieren. Den Ausgangspunkt der eigenen Untersuchungen zu diesem Problem bilden die von Ungeheuer entwikkelten Verfahren zur akustischen Sprecherklassifikation (Ungeheuer 1965). Deren mehr qualitative Ergebnisse werden anhand umfangreicheren Materials, das statistisch begründete Aussagen über die Signifikanz der Resultate erlaubt, durch apparativ quantifizierbare Kenngrößen ersetzt.

Das Untersuchungsmaterial besteht aus vier deutschen Texten, die von 20 männlichen Versuchspersonen (ein Text je zweimal) auf Tonband gesprochen werden; insgesamt stehen somit 100 Sprachaufnahmen zur Verfügung, die in ihrer Dauer zwischen 50 und 90 sec liegen. Aufgrund von Voruntersuchungen ist zu vermuten, daß Textstücke ab ca. 25 sec Länge eine 'textunabhängige' Sprecherunterscheidung ermöglichen, weil in diesem Fall die extrahierten Merkmale den jeweiligen Sprecher unabhängig von der Feinstruktur des gesprochenen Textes charakterisieren. Eine mithilfe der Informationsstatistik (Minimum discrimination information statistic, Kullback 1959) durchgeführte Untersuchung der phonologischen Transkriptionen der verwendeten Texte zeigt, daß diese bezüglich der Lauthäufigkeiten homogen sind. Deswegen können bei der akustischen Analyse zu Tage tretende Inhomogenitäten auf Sprechercharakteristika zurückgeführt werden.

Die Integrale über die Nulldurchgangsdichtefunktion $\rho_o(t)$ und die Extremwertdichtefunktion $\rho'_o(t)$ des Sprachsignals $s(t)$

$$R(\tau) = \int_o^\tau \rho_o(t)dt \qquad \text{und} \qquad R'(\tau) = \int_o^\tau \rho'_o(t)dt$$

stellen die ersten beiden Signalparameter dar, wobei die Integrationszeit τ gleich der Gesamtdauer von $s(t)$ gewählt oder apparativ fixiert werden kann. Das gibt die Möglichkeit, die Meßgrößen R und R' linguistisch oder zeitlich zu normieren — im ersten Fall je nach Sprechtempo unterschiedliche Analysezeiten für den gleichen

Text, im zweiten Fall konstante Analysierzeit τ, damit aber Realisierung unterschiedlich vieler Texteinheiten. Als weitere Parameter werden die Signalzeit T_S (Sprechzeit abzüglich aller Pausen), die Zeitdauer der stimmhaften Schallanteile T_V sowie die Zahl K_S der Unterbrechungen des Redeflusses bzw. K_V des Flusses stimmhaften Schalls bestimmt.

Die in Form zweidimensionaler Kontingenztafeln (Klassifikation der Messungen je eines Parameters nach Sprechern und Sprachaufnahmen) angeordneten Meßdaten (insgesamt 2400, durchschnittlicher Meßfehler von 1%) sind, wie der W-Test von Shapiro und Wilk (1965) bei einer Irrtumswahrscheinlichkeit α von 10% ergibt, in der Regel nicht normalverteilt.[1] Deshalb müssen zur Entscheidung der Frage, ob die akustischen Parameter bezüglich der einzelnen Sprecher signifikant verschieden sind, nichtparametrische Tests eingesetzt werden. Die Rangvarianzanalyse nach Friedman (1937) zeigt für die Sprecher-Stichproben aller Kontingenztabellen hochsignifikante Inhomogenitäten ($\alpha < 1\%$), d.h. deutliche Sprecherunterschiede sind vorhanden. Zur Lokalisierung dieser Unterschiede dient der Paarvergleich nach Wilcoxon (Wilcoxon matched pairs signed rank test) (Wilcoxon and Wilcox 1964) der zwar ein ziemlich aufwendiges Verfahren darstellt — pro Kontingenztafel ergeben sich bei 20 Sprechern $\binom{20}{2} = 190$ Einzeltests —, andererseits aber einen hohen Wirkungsgrad aufweist. Auf dem 10%-Niveau gesicherte Unterschiede sollen als signifikant gelten.

Die Resultate der statistischen Analyse sind als 'Klassifikationsmatrizen' dargestellt. Ein Zeichen im Feld X_{ij} zeigt an, daß die Sprecher i und j nicht zu unterscheiden sind. Da der Vergleich kommutativ ist und die Diagonalfelder X_{ii} stets besetzt sind, kann man sich auf die durch diese Diagonale begrenzten Dreiecksmatrizen beschränken. Abb. 1 zeigt hierfür ein Beispiel. Die obere Dreiecksmatrix gibt die Zerlegung der Sprechermenge aufgrund des Parameters R (gemessen an drei Aufnahmen über jeweils ca. 59 sec) wieder. Es gibt $z = 17$ Paare nicht zu unterscheidender Sprecher. Die 'einelementigen Klassen' *EK* sind wie die 'Restklassen' *RK*, die aus mindestens zwei Elementen bestehen und von den übrigen Untermengen der Zerlegung verschieden sind, explizit aufgeführt. Die untere Dreiecksmatrix ist das Ergebnis der Klassifikation aufgrund des entsprechenden Parameters R'. Man sieht, daß — wie sich generell bestätigt — R diskriminationsstärker als R' ist. Ferner sind die Klassifikationen anhand verschiedener Parameter nicht konsistent, d.h. die diskriminationsstärkeren Kennwerte liefern nicht nur feinere Zerlegungen, sondern auch solche mit anderen Klassenzugehörigkeiten der Sprecher.

Das gibt die Möglichkeit, durch Kombination mehrerer Parameter eine bessere Zerlegung zu erzielen. In Abb. 2 sind T_S und T_V zusammengefaßt: Bei den längeren Sprachaufnahmen (oben) sind bis auf 2 alle Sprecher voneinander zu unterscheiden, bei den kürzeren (unten) immerhin 13 Klassen (10 EK und 3 RK) zu bilden.

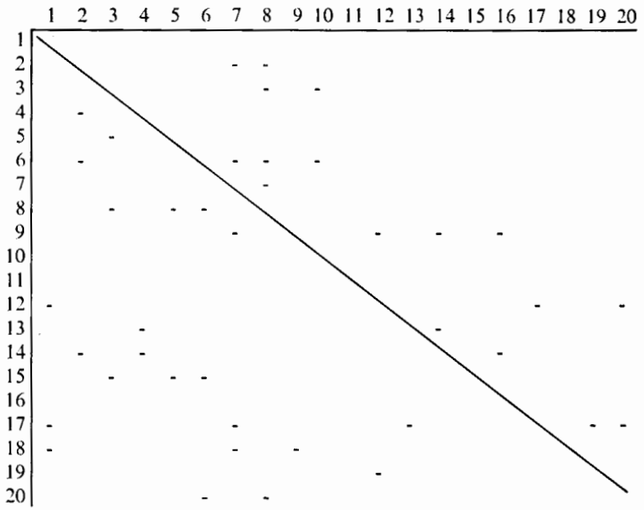
Zusammenfassend bleibt an Ergebnissen der Untersuchung festzuhalten:[2]

[1] Sämtliche statistischen Tests wurden unter Einsatz eines Computers (IBM 7090) durchgeführt.

[2] Eine detaillierte Darlegung der Ergebnisse und Beschreibung der elektronischen Meßschaltungen findet sich in Kirstein (1971).

R_2 (A3-A5)

z = 17
EK: (1) (4) (5) (11) (15) (18)
RK: (9,13,14,16) (12,17,19,20)
(2,3,6,7,8,10)



$R_{\frac{1}{2}}$ (A3-A5)

z = 25
EK: (10) (11) (16)
RK: (2,4,13,14) (3,5,6,8,15,20)
(1,7,9,12,17,18,19)

Abb. 1

(1) Jeder der 6 untersuchten akustischen Parameter ermöglicht eine Zerlegung der Sprechermenge in Klassen.

(2) Anhand der vorliegenden Daten läßt sich nicht entscheiden, ob die linguistisch oder die zeitlich normierten Kenngrößen R und R' diskriminationsstärker sind.

(3) Die Parameter T_S, T_V, K_S und K_V erweisen sich gegenüber R und R' als diskriminationsstärker und sind zudem mit geringerem technischem Aufwand am Sprachsignal zu messen.

(4) Mit wachsender Analysedauer wird die Zerlegung feiner.

(5) Durch Kombination mehrerer simultan meßbarer Parameter wird die Klassifikation besser; auf diese Weise kann der Verlust an Diskriminationsstärke bei kurzen Texten oder bei Reduzierung der Irrtumswahrscheinlichkeit kompensiert werden. Nimmt man alle 6 Parameter zusammen, so gelingt es sogar auf dem 1%-Niveau, 18 der 20 Sprecher voneinander zu separieren.

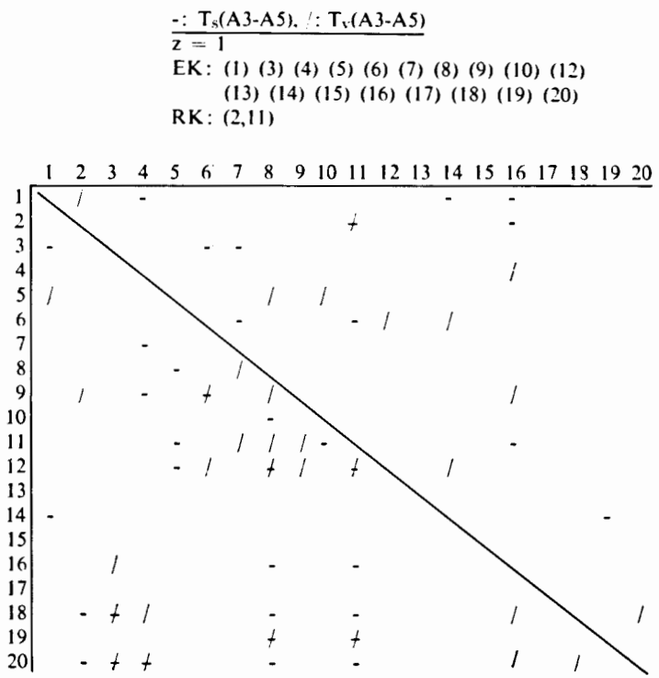


Abb. 2

(6) Die Diskriminationsstärke eines Parameters zeigt keine Abhängigkeit vom jeweiligen Text, wenn das Signal nur genügend lang (ca. 25 sec) analysiert wird. In diesem Sinn erlauben die untersuchten Parameter eine textunabhängige Klassifikation.

Institut für Kommunikationsforschung und Phonetik
Universität Bonn

LITERATURVERZEICHNIS

Friedman, M.
1937 "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance", *Journal of the American Statistical Association* 32:675-701.
Kirstein, M.
1971 *Akustische Untersuchungen zur automatischen Sprecherklassifikation* (H. Buske Verlag, Hamburg).

Kullback, S.
1959 *Information Theory and Statistics* (New York).
Shapiro, S.S. and M.B. Wilk
1965 "An Analysis of Variance Test for Normality (Complete Samples)", *Biometrika* 52: 591-611.
Ungeheuer, G.
1965 "Ein einfaches Verfahren zur akustischen Klassifikation von Sprechern", (Paper A 17, 5th International Congress on Acoustics, Lüttich).
Wilcoxon, F. and R.A. Wilcox
1964 *Some Rapid Approximate Statistical Procedures* (Lederle Laboratories, Pearl River, New York).

DISCUSSION

IIVONEN (Oulu)
Wurden die Parameter völlig automatisch gemessen oder waren auch manuell durchgeführte Messungen nötig?

KIRSTEIN
Die 6 Parameter wurden nicht manuell bestimmt, sondern apparativ gemessen.