ÜBER DIE GESTALTKRITERIEN AKUSTISCHER SPRACHSIGNALE

WALTER TSCHESCHNER*

Bei der Untersuchung der akustischen Zeichenerkennung durch organische Systeme, insbesondere der Sprachsignale, erhebt sich die Frage, welche Signaleigenschaften werden dem psycho-physischen Erkennungsvorgang zugrunde gelegt.

Beschränkt man sich auf die akustischen Signaleigenschaften, die die für die Artikulationseinstellung des organischen Senders typische Sendefunktion beschreiben, so reichen die Eigenwerte des Resonanztraktes bzw. die Formantdaten aus, um die semantische Funktion eindeutig zu erfassen. Die Formanten[1] können aus der Konfiguration des Ansatzrohres, also des Sendesystems, genau berechnet werden (1). Ihre exaktere Erfassung aus dem Sendespektrum gelingt wegen dem diskontinuierlichen Anregungsspektrum bei stimmhaften Lauten nur über Näherungsverfahren (2). Die besondere Bedeutung der Formanten ist dadurch gekennzeichnet, daß bei ihrer hinreichend genauen Bestimmung nach Frequenz, Bandbreite und Amplitude drei Formanten genügen, die Zuordnung Zeichenträger — elementares Sprachzeichen (Artikulationsstellung) zu finden.

Für den Entwurf technischer Spracherkennungs- und Verarbeitungseinrichtungen ist es nun von großer Bedeutung zu wissen, welche Kenngrößen der adäquate organische Empfänger aus dem Signalkurzzeitspektrum extrahiert und der Erkennung zugrunde legt. Wie bereits früher mit Hilfe der subjektiven Analyse — durchgeführte Untersuchungen zeigten, stützt sich der organische Empfänger nur mittelbar auf die Extrahierung von Formantdaten (3). Zur Erfassung der primären empfangsseitigen Kenndaten von Lautspektren wurden in weiteren Versuchen u. a auch synthetische Lautspektren verwendet.

Über die dabei gewonnenen Ergebnisse soll hier berichtet werden.

Zu den Versuchen wurde ein handgesteuerter Synthesator (Bild 1) benutzt, der nach dem Prinzip der Sprachdecodierung eines Kanalvocoders (4) arbeitet. Das System besitzt 12 parallel geschaltete Netzwerke, deren Durchlaßbereiche und Frequenzstaffelung etwa der Frequenzgruppenbildung entsprechen:

---

* VEB Elektronische Rechenmaschinen.

[1] Nach Fant (1) ist es zweckmäßig, den Terminus Formant nur auf die Dämpfungsnullstellen des Ansatzrohres anzuwenden.

0 — 0,25 — 0,35 — 0,45 — 0,6 — 0,85 — 1,15 — 1,45 — 1,75 — 2,1 — 2,5 — — 2,95 — 3,5 kHz

Das Spektrum jedes Reizes wurde über Steckbuchsen der steuernden Widerstandsmatrix (Bild 1) vorgegeben. Die Auslösung wurde von Hand vorgenommen, wobei ein Zeitfunktionsgenerator mit einer kontinuierlich verlaufenden Zeithülle von 200 ms Länge freigegeben wurde.

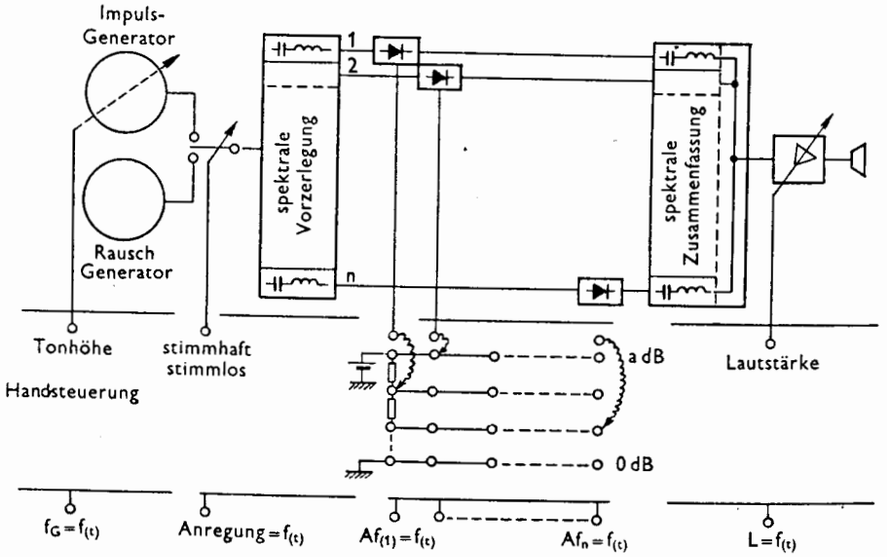

Abb. 1. Schema der Sprachsignalsynthese.

Als Anregungsparameter wurden verwendet:

Stimmhaft: Dreieckfunktion — 6 $dB/_{\text{Oktave}}$ ausgeglichen auf 0 $dB/_{\text{Okt}}$, f_G = = 110 Hz, Grundfrequenzmodulation zur Erhöhung der Natürlichkeit $f_{G\,\text{Hub}} = +20$ Hz, $f_{\text{mod}} = 0{,}2$ Hz

Stimmlos: Weißes Rauschen 0 $dB/_{\text{Oktave}}$.

Je 50 Reize, von denen jeder 4 bis 5mal vorkam, wurden in einer bestimmten zeitlichen Reihenfolge auf Band zusammengestellt und über dynamische Meßhörer bei einer mittleren Lautstärke von 80 phon von 5 Hörern beurteilt. Nach Vorgabe eines nichtbewerteten Versuches konnte jeder Reiz, sofern akzeptabel, mit einem geläufigen Lautnamen belegt werden. Es ergab sich jedoch, daß die Zahl der nicht treffbaren Zuordnungen außerordentlich gering waren.

Im 1. Versuch wurde nur jeweils ein Kanal bei stimmhafter Anregung geöffnet. Dabei wurde darauf geachtet, daß die zeitliche Aufeinanderfolge ähnlicher Reize vermieden wurde. Beim 2. Versuch wurden die gleichen Einstellungen mit stimmloser Anregung wiederholt. Im 3. Versuch wurden bei jedem Reiz der in Bild 2 markierte

Kanal mit voller Amplitude geöffnet und entsprechend dem 1. Versuch ein weiterer Kanal bei stimmhafter Anregung zugeschaltet.

Das Ergebnis der Untersuchung in Bild 2 enthält den Prozentsatz der pro Reiz getroffenen Zuordnungen (5).

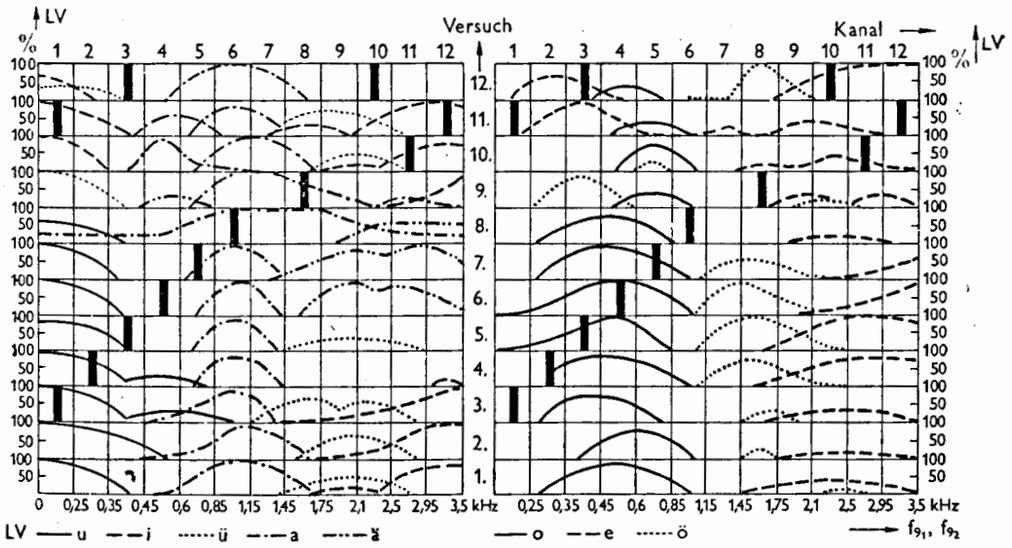

Abb. 2. Die Lautverständlichkeit synthetischer Spektralkonfigurationen.

Wird die Formanthypothese zur Deutung der Versuchsergebnisse herangezogen, dann müßten sich die Hörerurteile nach der folgenden Formanttabelle für deutsche Vokale richten.

	F_{1kHz}	F_{2kHz}	F_{3kHz}
u	0,25	0,6	
o	0,4	0,8	
ä	0,6	1,2	2,2
a	0,85	1,2	2,5
ö	0,4	1,8	2,4
e	0,4	2,4	3,25
ü	0,35	2	2,8
i	0,25	2,6	3,5

Demzufolge dürften im 3. Versuch nur die Lautvalenzen [*u*], [*ü*] und [*i*] im 5. Versuch nur [*o*], [*ö*] und [*e*] und im 6. Versuch nur [*u*] und [*ä*] in Erscheinung treten. Nach Bild 2 zeigen sich jedoch andere Ergebnisse.

Im 3. Versuch werden z. B. teils unsichere [*ä*] 5 %, [*ö*] 25 %, [*e*] 30 %, teils sichere [*ü*] 65 %, [*o*] 80 %, teils sehr sichere [*a*] 90 %, [*i*] 95 % [*u*] 100 % Zuordnungen gefunden. Diese Werte stellen Maximalwerte dar, die rechts und links davon auf der Kanalskala nach Null verlaufen. Entsprechende Verhältnisse sind in allen Versuchen zu beobachten.

In der gewählten Versuchsanordnung, bei der den Hörern Relativentscheidungen möglich sind, werden sehr sichere Vokal-Zuordnungen (> 90 %) getroffen, wenn folgende Kanäle geöffnet werden.

Kanal

u	1	2+1 4+1 5+1	1
o	4	3+4 4+3 5+3	4
ä	4+9	5+11	5+10
a	6	1+6 2+6 3+6 4+6 5+6 6+7 6+8 8+6 8+7 11+7 3+10+6	6
ö	3+8	3+10+8	3+8
e	3+11	3+10+11 1+12+3 3+10+12	3+11
ü	8+1		1+8
i	12	1+12 11+1 1+12+11	1+11/12

Es ist zu erkennen, daß zwar für die Kanalmuster, die näherungsweise der oben angegebenen Formantkonfiguration entsprechen, weitgehend die zugehörigen Lautnamen gefunden werden, für die verschiedensten anderen Kanalmuster werden aber ebenfalls sehr sichere Zuordnungen angegeben. Somit kann das Formantprinzip nicht der Erkennung zugrundeliegen. Wird dagegen hier zur Beschreibung des Erkennungsprozesses die Gestalt und nicht einzelne diskrete Punkte der Spektralhülle herangezogen, so kann der Versuch befriedigend gedeutet werden.

Sieht man von den zusätzlichen Einflußfaktoren, wie Adaption, Folgekontrast usw. ab, so können aus diesem Versuch, ergänzt durch Einzelversuche, etwa die in Bild 3 enthaltenen Gestaltkonfigurationen der Spektralhülle der Erkennung deutscher Vokale zugrundegelegt werden. Diese Figuren, die man sich als Transformation der Auslenkung der Basilarmembran vorzustellen hat, werden durch kettenlinienförmige Merkmalschwellen begrenzt. Als typisch sind die Formmerkmale und erst in zweiter Linie die Frequenzlage anzusehen. Formmerkmale sind z. B.:

[*u*] Frequenzgruppenbreite Erregung niederfrequenter Lage. Besonderes Merkmal niederfrequente Anteile unter dem Maxima, mindestens um 40 dB gedämpft.

[*a*] Zentrale maximale Erregung, ca. 3 Frequenzgruppen breit.

[*i*] Ausgeprägtes Minimum, dessen Sohle mindestens 25 dB unter den Maxima liegen muß.

Ist ein solches Formmerkmal Bestandteil eines Kanalmusters, dann wird der entsprechende Laut zugeordnet, wobei offenbar in der Erkennungslogik noch eine hierarchische Rangordnung wirksam wird (6).

Somit kann einem akustischen Reiz analog einem optischen Reiz eine Gestaltkonfiguration und eine Färbung (Grundfrequenz) zugeordnet werden.

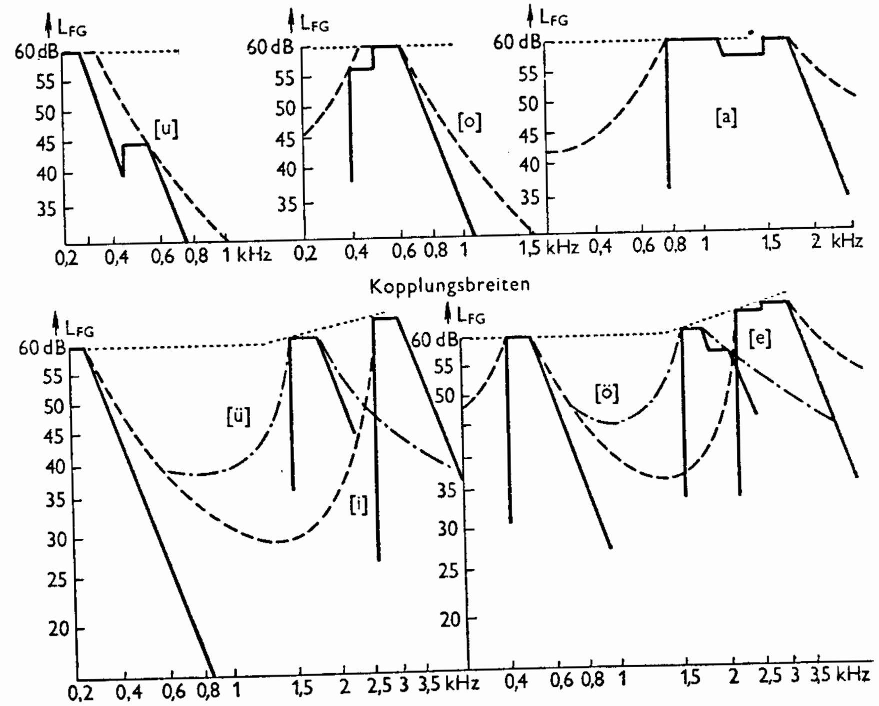

Abb. 3. Die zu einer Lautvalenzklasse eines deutschen Vokales gehörige empfundene Gestaltfunktion ermittelt durch Syntheseuntersuchungen.

——— dominierendes Gestaltmerkmal.

- - - - - - Merkmalschwelle.

. . . . . . Größe der spezifischen Lautheit bei konstantem Frequenzpegel.

LITERATUR

1. Fant, G.: *Acoustic theory of speech production.* Verlag Mouton Co. 1960, 's-Gravenhage.
2. Lindblom, B.: Accuracy and limitation of sonagraph—measurements. *Proc. IV. Phonetik Kongress* Helsinki 1961, S. 188—200.
3. Tscheschner, W.: Analyse der deutschen Sprache unter besonderer Berücksichtigung der nichtstationären Vorgänge (Teil 4). *Zeitschr. f. Phonetik und Kommunikationswissensch.* 19 (1966), S. 141—201.
4. Krocker, E.: Aufbau und Untersuchung eines Übertragungssystems für synthetische Sprache. *Wiss. Zeitschr. der TH.* Dresden 6 (1956/1957) H. 4, S. 757—776.
5. Tscheschner, W.: Die Sprache und ihre Erkennung. Habilschrift (unveröffentl.). TU Dresden 1967.
6. Tscheschner, W.: Ein Beitrag zur subjektiven Erkennung akustischer Sprachlaute. *IV. Ak. Konferenz* Budapest 1967 (erscheint im Tagungsbericht).

DISCUSSION

Fant:

Your results appear to support a two-formant model of speech perception in the sense of two major spectral stimulus regions.