Verh. 5. int. Kongr. Phon. Wiss., Münster 1964, pp. 582–588
(S. Karger, Basel/New York 1965).

Perzeptive Grenzen der Phonem-Unterscheidung

Von F. Winckel, Berlin

Während man anfänglich von lautanalytischen Untersuchungen ausging (Spektralanalyse), standen in den letzten Jahren die Untersuchungen mittels Segmentation im Vordergrund, die zu Klärungen führten durch die Bildung von *perceptual segments*[1]. Wichtig war die Ermittlung minimaler Phonem-Dauern, die für vokalische Laute 10 bis 15 ms beträgt[2]. Es reichen zwei Perioden zum Erkennen eines Lautes aus. Schließlich ist die digitale Synthese von Sprachlauten (Computer) gelungen[3]. Daraus geht der Aufwand an Elementarquanten für die Phonembildung hervor.

Die Problematik der Untersuchung besteht darin, daß das Phonem kein stationäres Gebilde ist, sondern in der kurzen Dauer der Intonation Änderungen in der Grundfrequenz wie auch der spektralen Zusammensetzung unterworfen ist. Die Frage ist, in welchen Grenzen solche Änderungen vom Ohr bemerkt werden. Im Sprachzusammenhang wird das komplex zusammengesetzte und zeitlich veränderliche Phonem als Ganzheit wahrgenommen. Die phonemische Folge überträgt nicht soviel Sprach-Information wie das Oszillogramm des Sprachschalls.

Nimmt man für normale Sprechgeschwindigkeit in der englischen Sprache einen Vorrat von 40 Phonemen an[4], so erhält man 5,3 bit pro Phonem, wobei strukturelle Merkmale ausgelassen sind, und mit einer sehr schnellen Folge von 10 Phonemen/sec den Wert 53 bit/sec. Hierbei kann es sich jedoch nur um einen theoretischen Wert handeln, denn die Vokale schrumpfen bei hoher Sprechgeschwindigkeit zusammen zu dem neutralen Laut ∃[6] (Abb. 1)[9].

Die Grenzen der Sprachlauterkennung sollen im folgenden aus der Rasterschärfe der Hörfläche ermittelt werden. Tragen wir für jeden Punkt der Hörfläche das Unterscheidungsvermögen für Frequenzen und Lautstärken bei Darbietung von Sinustönen ein, so erhalten wir eine Rasteraufteilung nach *E. Zwicker* gemäß Abbildung 2. Die hörbare Änderung der Intensität △ I und der Frequenz △ f wurde stufenweise über die Hörfläche bestimmt durch Ermittlung der Amplitudenmodulations- und der Frequenzmodulationsschwellen. Es wurden Modulationsfrequenzen von 4 Hz gewählt, weil das Ohr in diesem Bereich die größte Empfindlichkeit besitzt. Für die Untersuchung von Sprache erscheint die Wahl günstig, soweit es die Intensität betrifft, da bei der Sprechgeschwindigkeit von

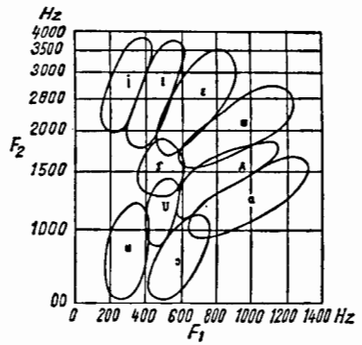

Abb. 1. Die Allophonklassen amerikanischer Vokalphoneme nach Messungen von *Peterson* und *Barney* (entnommen aus *W. Meyer-Eppler*, Informationstheorie, Berlin 1960).

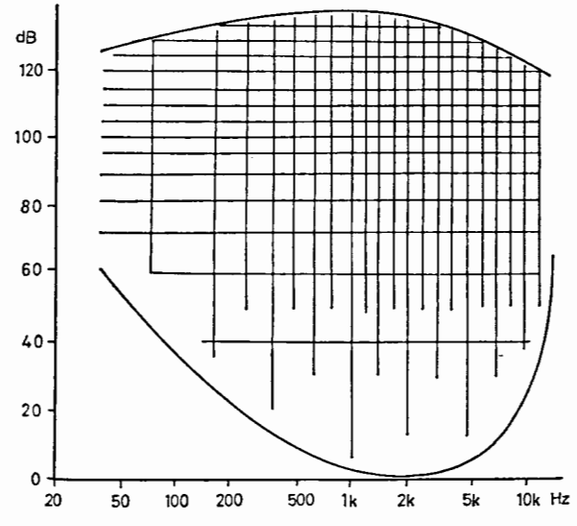

Abb. 2. Unterscheidbare Abschnitte der Hörfläche bei Sinustönen > 2 sec (je 1000 zusammengefaßt).

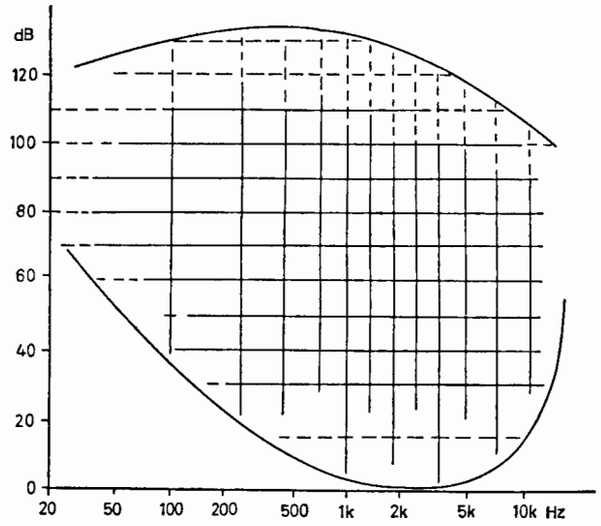

Abb. 3. Unterscheidbare Abschnitte der Hörfläche bei Breitbandrauschen (je 100 zusammengefaßt).
Muster für Reibe- und Zischlaute.

vier Silben/sec ein häufig vorkommendes Anschwellen der Sprechintensität in diesem Rhythmus erfolgt.

Dieselben Versuche wurden wiederholt bei Darbeitungen von Rauschen (Abb. 3) anstelle von Sinustönen. Aus den Messungen[11] erhält man 132 unterscheidbare Tonhöhenstufen und 120 hörbare Intensitätsunterschiede im Bereich von 10–130 dB. Somit ergeben sich aus dem Hörvergleich über die Hörfläche 15840 Valenzen von Geräuschen, dagegen 300000 von Sinustönen. Streng periodische Sinustöne kommen in der Sprache gar nicht vor, in der Musik nur selten. Vielmehr ist ein Geräuschcharakter vorherrschend (Abb. 4), so daß das dem Geräusch entsprechende Raster eher den praktischen Verhältnissen entspricht. Die Zahl der Valenzen entspricht in diesem Fall nur 5% gegenüber den von Sinustönen. Somit werden Intensitätsschwankungen von weniger als 1 dB und Tonhöhenschwankungen von $\pm$ 10 Hz bei Rauschdarbietung nicht bemerkt.

Nun gelten die angegebenen Schwellwerte für Intensitätsänderung nur für Impulsdauern oberhalb 0,1 sec. Darunter findet ein Schwellenanstieg statt[11]. Nimmt man einen noch möglichen Wert von 10 ms für das Kurzphänomen an, so gelangt man zu einem Intensitätszuwachs der Schwelle von 2 dB. Andererseits gelten die 250 unterscheidbaren Tonstufen über die ganze Frequenzskala nur

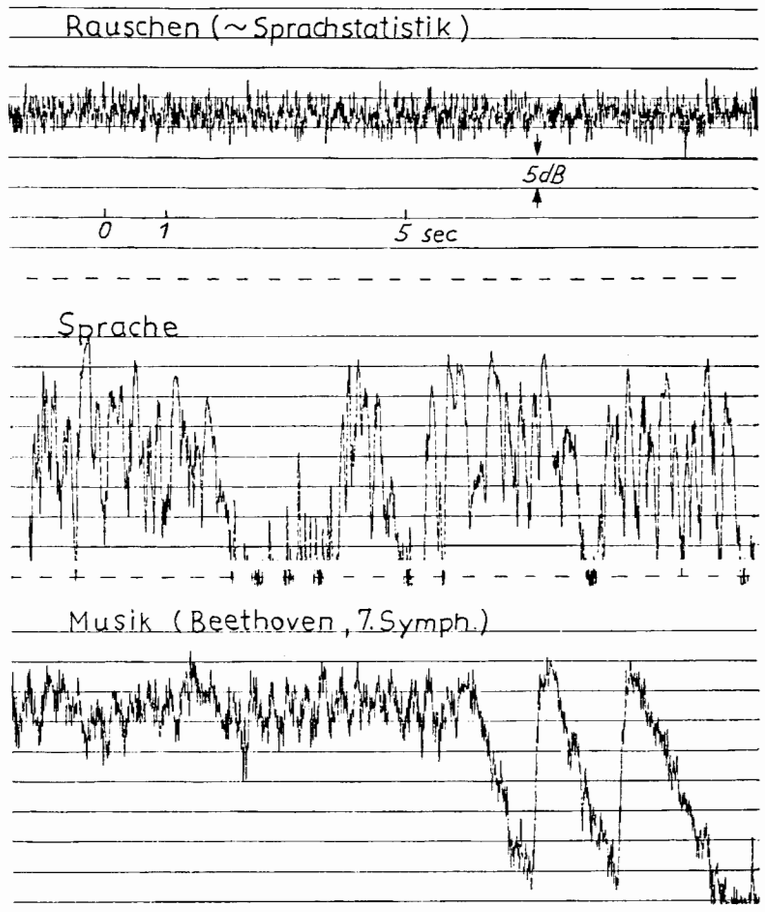

Abb. 4. Der Rauschcharakter von Sprache und Musik im Vergleich zu weißem Rauschen, das gemäß der Sprachstatistik gefiltert wurde.

für Dauertöne. Bei Darbietung von Tonimpulsen von etwa 10 ms unterscheidet das Ohr im ganzen nur 120 Tonstufen anstelle der ganzen 850 Stufen bei einer Dauer von über 250 ms[3]. Das Raster für Sinustöne muß daher 1:7 in der Frequenzskala und 1:2 in der Intensitätsskala geteilt werden, um Kurztönen von 10 ms zu entsprechen (Abb. 5). Mit abnehmender Dauer geht die Tonalität in einen geräuschähnlichen Laut über, was bei 10 ms bereits gegeben ist. Dies ist der Bereich der Explosivlaute. Andererseits bedarf die Vokalerkennung einer Mindestdauer von etwa 50 ms. Bei schnellem Sprechen werden Vokaldauern von 100 ms nur unwesentlich über-

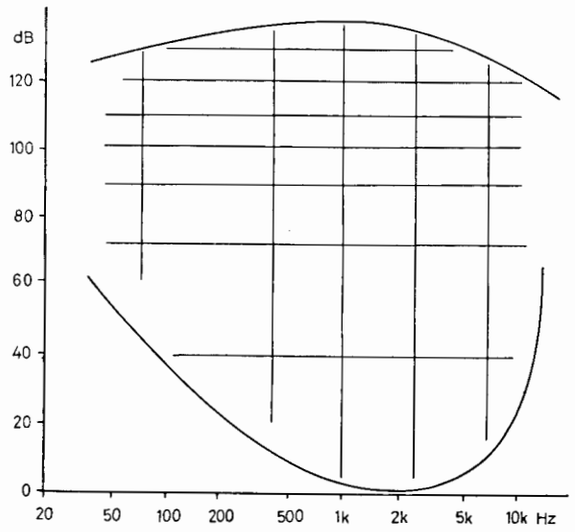

Abb. 5. Unterscheidbare Abschnitte der Hörfläche bei Sinustönen von 10 ms Dauer (je 1000 zusammengefaßt).
Muster für Explosivlaute.

schritten. Für diesen Wert gibt *Feldtkeller* bei Sinuston-Einwirkung 400 statt 850 erkennbare Tonstufen an. Das Raster für Dauertöne

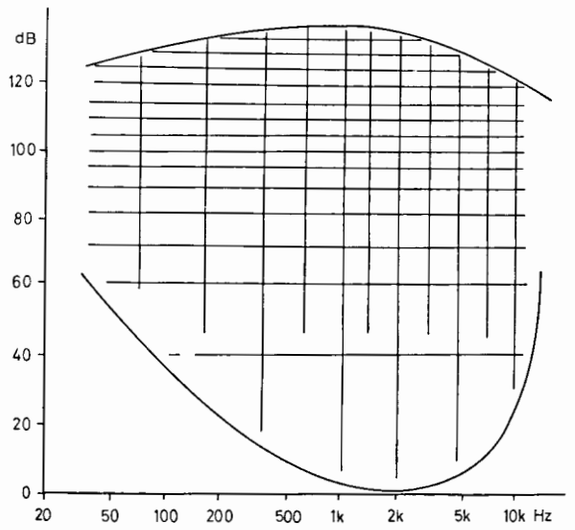

Abb. 6. Unterscheidbare Abschnitte der Hörfläche bei Sinustönen von 100 ms Dauer (je 1000 zusammengefaßt).
Muster für Vokale.

muß also in der Anwendung auf Vokale in der Frequenzskala 1:2 reduziert werden (Abb. 6). Bei strenger Vokalperiodizität der Frequenz würde eine Verschärfung der Tonhöhen-Empfindung durch Bildung des Residuums[7] eintreten. Die tatsächlich vorhandene Schwankung der Vokalfrequenzen macht die Tonhöhe-Empfindung durch das Residuum wieder unscharf.

Für die Berechnung des Informationsflusses von Sprache ist zu berücksichtigen, daß dafür nur ein kleiner Anteil der Valenzkapazität über die Hörfläche anzusetzen ist, nämlich für das angenommene ersatzweise Rauschen weniger als ein Viertel[10]. Mit einem Vorrat von nur 3400 Valenzen und 10 Phonemen/sec ergibt sich der Informationsfluß zu 115 bit/s, für den Vorrat aller Elementarlaute der Musik zu 125 bit/sec. Die Reduktion auf Telephonsprache ergibt 10,8 bit.

Die Untersuchung der Unterschiedsschwellen, die auf eine beträchtliche Unempfindlichkeit des Ohres in dieser Beziehung schließen lassen, geben den Hinweis, wie wenig die Absolutempfindung der Tonhöhe, Lautstärke, Klangfarbe, Dauer beim Erkennen von Sprache Bedeutung hat. Scheinbar sind daran noch andere Faktoren beteiligt, die sich möglicherweise in spezifischen Zusammenschaltungen des Zentralnervensystems auswirken. *W. D. Keidel* gibt den Hinweis, daß in der sekundären Rinde ganze Frequenzgruppen, wie sie für die Sprachlaute gebraucht werden, schon zusammengeschaltet sein könnten. Auch in der primären Rinde zeigt sich ein spezifisches Verhalten in der Frequenzskala, das nicht mit der Frequenzverteilung auf der Basilarmembran übereinstimmt[5].

Literatur

1. *Cohen, A.* and *'t Hart, J.:* Segmentation of the speech continuum. 4th Int. Congress Acoustics (G 51) (Kopenhagen 1962).
2. *Cramer, B.:* Über das Erkennen der Sprachlaute, in Aufnahme und Verarbeitung von Nachrichten durch Organismen (Vorträge NTG-Karlsruhe) (Stuttgart 1961).
3. *Feldtkeller, R.:* Wechselbeziehungen zwischen Psychologie, Physiologie und Nachrichtentechnik, in Aufnahme und Verarbeitung von Nachrichten durch Organismen (Vorträge NTG-Karlsruhe) (Stuttgart 1961).
4. *Fry, D. B.* and *Denes, P.:* The role of acoustics in phonetic studies in: *Richardson, E. G.* and *Meyer, E.:* Sound (Amsterdam 1962).
5. *Keidel, W. D.:* Ergebnisse der Elektrophysiologie des Hörens in: *Schubert, K.:* Theorie und Praxis der Hörgeräteanpassung (Stuttgart 1960).
6. *Lindblom, B.:* On vowel reduction. Diss. (Uppsala 1963).
7. *Schouten, I. F.:* Proc. Acad. Amsterdam *43:* 991 (1940).
8. *Strasser, B. E.* and *Matthews, M. V.:* Music from Mathematics, Bell Telephone Lab. 1961 (Beiheft zu einer Schallplatte).

9. *Ungeheuer, G.:* Anwendung der Störungsrechnung in der Eigenwerttheorie der Vokalartikulation, S. 238, 3rd Int. Congress Acoustics, Stuttgart 1959 (Amsterdam 1961).
10. *Winckel, F.:* Das Gehör in informationstheoretischer Behandlung. Arch. Ohr.-Nas.-KehlkHeilk. *182:* 456–470 (1963).
11. *Zwicker, E.:* Informationskapazität des Gehörs. Acustica *6:* 365–381 (1956).

Adresse des Autors: Prof. Dr.-Ing. F. Winckel, Technische Universität, Hardenbergstraße 34, *1 Berlin 12* (Deutschland).

Discussion

von Essen (Hamburg): Durch die Untersuchungsergebnisse ist bewiesen, daß Silbentonhöhen durch Abhören – entgegen der Behauptung von *Peters* – doch bestimmbar sind. Im übrigen wird damit erklärt, daß Gleittöne besonders in langen Silbenträgern, deren Dauer über die 0,2-Grenze hinausgeht, apperzipiert werden und dann auch phonologische Relevanz erlangen können.

Rothauser (Wien): Bei Versuchen mit synthetischer Sprache haben wir festgestellt, daß eine Rauschmodulation des Grundtons allein bereits bei einer mittleren Änderung von $\sim 3\,^0/_{00}$ der ursprünglichen Grundfrequenz als Rauhigkeit des Sprachklanges bemerkbar wird. Dies scheint im Gegensatz zu dem reduzierten Tonhöhenunterscheidungsvermögen zu stehen, das vom Referenten beschrieben wird.

Janota (Prag): 1. Das Ergebnis des Experimentes mit dem Posaunengleitton stimmt gut mit der alten Erfahrung der Phonetiker überein, daß auch die Musikologen und Musiker mit einer genauen Bestimmung der Tonhöhe in zusammenhängender Rede Schwierigkeiten haben. Der Dauer des Lautes nach hören sie die gleitende Tonhöhenveränderung als eine konstante Stufe oder aber als eine schwer zu bestimmende Tonhöhe.

2. Wenn wir nun einen Gleitton zerschneiden, dann empfindet das Ohr diese Abschnitte des Gleittones als konstante Stufen. Damit ist aber keineswegs bewiesen, ob das Ohr beim Beurteilen der Sprache auf Tonhöhen*verlauf* oder Tonhöhen*stufen* reagiert.