Zeitschrift für experimentelle und angewandte Psychologie
1987, Band XXXIV, Heft 1, S. 64—71

Manuelle versus elektronische Analyse von Sprechpausen

Thomas Klos und Heiner Ellgring

Zentralinstitut für Seelische Gesundheit, Mannheim, und Max-Planck-Institut für Psychiatrie, München

Es wird gezeigt, daß die digitale Sprachanalyse bei der Messung von Sprechpausen unter bestimmten Umständen Fehler aufweist. Am Beispiel von 16 standardisierten Interviews mit depressiven Patienten wurden Sprechpausen von Patienten und Interviewern nach zwei Methoden gemessen: mit einer einfachen manuellen Methode deren Interraterreliabilität bei .88 und höher lag und nach Methoden der digitalen Sprachverarbeitung. Die Ergebnisse beider Analysen wurden verglichen. Dabei zeigte sich, daß die manuelle Methode für Sprechpausen oberhalb 390 ms reliabel ist und gleiche oder bessere Ergebnisse bringt. Bei qualitativ schlechten Tonaufnahmen ist diese manuelle Methode vorteilhaft.

1. Einleitung

Die Analyse von Sprechpausen spielt in der Psycholinguistik, Phonetik und Sprachwissenschaft eine bedeutende Rolle. Auch in der klinischen Forschung wurden Sprechpausen in der Sprache von Schizophrenen (Clemmer, 1980) Depressiven (z.B. Greden & Carroll, 1980; Greden et al., 1981) und bei Patienten mit Parkinsonismus (Mawdsley & Gamsu, 1971) untersucht.

Zur objektiven Messung von Sprechpausen wurden eine Reihe von Apparaturen entwickelt, die von halbautomatischer bis zu elektronischer Messung per Oszilloskop und Spektrograph reichen. Eine gute Übersicht hierzu gibt Kuenzel, 1974.

Die heute elaborierteste Methode zur Bestimmung von Sprech- und Artikulationspausen bietet die digitale Sprachanalyse (vergl. auch Scherer, 1982). In der folgenden Darstellung soll gezeigt werden, daß auch die digitale Sprachanalyse unter bestimmten Bedingungen fehlerhafte Daten liefert. Eine einfache manuelle Methode wird als alternative Möglichkeit zur Analyse von Sprechpausen vorgestellt.

2. Material

Im Rahmen einer Untersuchung von Sprechgeschwindigkeit und Sprechpausen bei depressiven Patienten (Klos & Ellgring, 1984) analysierten wir Sprachmaterial aus standardisierten Interviews zum Verlauf depressiver Erkrankungen (Ellgring et al., 1978). Die 32 Interviews stammten von 11 weiblichen und 5 männlichen Patienten, jeweils im depressiven und gebesserten Zustand. Alle Interviews fanden im Aufnahmestudio der psychologischen Abteilung des MPI München statt. Patient und Interviewer saßen in einem Winkel von 90 Grad zueinander. Die Tonaufnahme erfolgte mit zwei Krawattenmikrofonen. Die Gesamtdauer des standardisierten Teils der Interviews betrug fünf bis zehn Minuten (28 Fragen). Dann folgte ein freies Gespräch zwischen Interviewer und Patient von ca. fünf Minuten Dauer.

Zur Sprachanalyse wurden folgende Äußerungen herangezogen: Auf der Seite des Interviewers Frage 1, 2, 27 und 28 sowie eine längere Äußerung (größer 25 Silben) aus dem freien Teil. Auf der Seite des Patienten die entsprechenden Antworten. Die Äußerungen der Interviewer waren im Mittel 9.1 sec. lang, die der Patienten 29.5 sec. Das untersuchte Material war also heterogen aus spontanen, rezitierten, kurzen und langen Äußerungen zusammengesetzt.

3. Methode

Zunächst wurden die Sprachproben mit einer Abtastfrequenz von 16380 Hz und einer Filterfrequenz von 6553 Hz digitalisiert. Mit Hilfe eines Programms zur Pausendetektion (Verfahren nach Hess, vergl. Helfrich, 1980) wurden alle Stilleperioden mit einer Dauer von 390 ms und länger markiert. Die kurzen, rezitierten Fragen der Interviewer wurden mit einem Schwellenwert von 94 ms abgetastet, um auch den Bereich zwischen 100 ms bis 390 ms zu erfassen. Mit einem weiteren Programm wurden Anzahl und Dauer der Sprechpausen ausgegeben. Diese quantitative Analyse erlaubt keine Angabe über die Position von Stilleperioden im Kontext. Um die Ergebnisse des Rechners zu überprüfen wurde von 50% des Materials ein vokalisationsgetreues Transkript der Äußerungen erstellt (incl. gefüllten Pausen, Versprecher, Repetitionen, Affektlaute, Atemgeräuschen, Backchannelsignalen des Gesprächspartners, Bewegungsgeräuschen usw.). Mit Hilfe dieses Transkriptes und einem Programm zur opto-akustischen Segmentierung des Signals (siehe Standke, 1980) wurde für jede vom Rechner ermittelte Stilleperiode die genaue Position im Kontext markiert. Dabei fielen folgende Artefakte auf:

1. In einer Reihe von Stilleperioden tauchten trotz Verwendung von Krawattenmikrofonen Backchannelsignale („mhm") des Gesprächspartners

auf. Eine solche Stilleperiode wird bei der Pausendetektion zerlegt. Dies verfälscht die ausgegebene Anzahl und Dauer der Stilleperioden.

2. Ist das Sprachsignal schwach, durch die zum Teil leise Stimme des Patienten in depressivem Zustand, so erfolgt keine klare Trennung zwischen Signal und Rauschen. Läßt man vom Rechner markierte Stilleperioden aus solchen Passagen durch einen D/A-Wandler ausgeben, so enthalten diese Stilleperioden zu Beginn und/oder am Ende Signale und nicht nur das als Pausenkriterium ermittelte Hintergrundrauschen.

Mit folgender Methode sollte die Möglichkeit einer manuellen Analyse überprüft werden:

a) Die mit einer Bandgeschwindigkeit von 19 cm/sec. aufgezeichneten Äußerungen wurden mit halber Geschwindigkeit (9.5 cm/sec.) abgespielt. b) Ein Hochpaßfilter diente als Ausgleich für die mangelnde Verständlichkeit der Äußerungen. (Filterfrequenz je nach Stimmlage 100 bis 180 Hz). c) Ein Rater suchte per Kopfhörer jede Äußerung mit Konzentration auf Sprechpausen ab. d) Er markiert jede gefundene Stilleperiode in dem Transkript. e) In einem zweiten Durchgang ermittelt der Rater mit Hilfe des Transkriptes die Dauer der Stilleperioden mit einer empfindlichen Digitaluhr.

Am Beispiel von 6 zufällig ausgewählten Interviews wurde die Interraterreliabilität durch einen zweiten Rater geprüft. Die Werte lagen für die Markierung von Stilleperioden bei .90 (V_2 nach Holsti) bzw. .89 (Cohen's „Kappa"). Für die Zeitmessung von gemeinsam an gleicher Position markierten Stilleperioden ergab sich eine Korrelation von .98 bei n = 123 Stilleperioden. Die Mehrfachwahlreaktionszeiten der beiden Rater betrugen 542 und 552 ms. Entsprechende Interraterreliabilitäten wurden durch einen Rater mit höherer Reaktionszeit (672 ms) und anderem Geschlecht ermittelt. Es ergaben sich Werte von .89 (V_2 nach Holsti) und .88 (Cohen's „Kappa") für die Markierung von Stilleperioden und eine Korrelation von .96 (n = 72) für die Zeitmessung am Beispiel von 4 der sechs ausgewählten Interviews.

Die Interraterreliabilität war somit ausreichend, um einen Vergleich der Ergebnisse aus der elektronischen und der manuellen Analyse der Interviews durchzuführen. Bei diesem Vergleich wurden aus der manuellen Auswertung nur solche Pausenwerte genommen, die länger als 390 ms plus dem mittleren Meßfehler von 160 ms waren. Dies sollte verhindern, daß aus der manuellen Analyse Pausen eingingen die kürzer als 390 ms waren und somit vom Rechner bei einem Schwellenwert von 390 ms nicht erfaßt werden konnten. Entsprechend wurde mit den kürzeren Pausen aus den Standardfragen der Interviewer verfahren.

4. Ergebnisse

Beim Vergleich zwischen elektronisch und manuell ermittelten Stilleperioden können vier Kategorien von Sprechpausen unterschieden werden. Solche, die von Rechner und Rater an gleicher Position ermittelt wurden (G = gemeinsame) und elektronisch zusätzliche (EZ), die nur vom Rechner erfaßt, d.h. vom Rater nicht wahrgenommen wurden. Auditiv zusätzliche (AZ) sind Pausen, die nur vom Rater erfaßt wurden, vom Rechner jedoch keine Diskrimination zwischen Signal und Rauschen möglich war, z.B. bei zu leiser Stimme und/oder hohem Hintergrundrauschen. Schließlich gibt es noch Stilleperioden, die elektronisch fälschlicherweise gespalten wurden (FG) weil sie ein Backchannelsignal des Gesprächspartners („mhm") enthielten, oder — in Ausnahmefällen — ein Signal durch Schlucken oder Bewegungen auftrat.

Tabelle 1 zeigt für die 16 oder 32 Interviews für die eine manuelle und

Tabelle 1

Korrelationen zwischen auditiv und elektronisch gemessenen Stilleperioden für die einzelnen Interviews und Häufigkeit von gemeinsamen (G), elektronisch zusätzlichen (EZ), auditiv zusätzlichen (AZ) und fälschlich gespaltenen (FG) Stilleperioden. Z = Zustand des Patienten.

Nr/Z	R	G	EZ	AZ	FG
1/−	.979	24	1	3	2
2/+	.908	18	1	2	4
3/+	.975	22	11	2	1
4/−	.982	15	4	2	3
5/−	.998	20	4	1	0
6/+	.987	20	3	2	2
7/+	1)			2)	
8/−	.993	22	1	1	3
9/−	.978	31	1	0	0
10/+	.937	11	3	2	0
11/+	.909	24	7	2	0
12/−	.968	33	6	2	0
13/−	.980	22	2	2	2
14/+	.953	19	6	3	3
15/−	.995	30	6	7	3
16/+	.948	19	6	3	1
Ges.:	.982	330	62	34	24
%		77.5	14.6	7.9	

1) Interview wegen zu hohem Geräuschpegel elektronisch nicht auswertbar.
2) 18 Stilleperioden auditiv faßbar.

eine opto-akustisch überprüfte elektronische Auswertung vorlag, die entsprechenden Häufigkeiten in jeder der vier Kategorien.

Die zweite Spalte in Tab. 1 gibt die Korrelation zwischen den gemeinsamen Pausen an. In der ersten Spalte ist Interviewnummer und Zustand (Nr/Z) des Patienten angegeben. Ein Vergleich zwischen den Spalten eins und zwei zeigt, daß die Übereinstimmung zwischen gemeinsamen Pausen bei depressivem Zustand (−) des Patienten in der Regel etwas höher ist, als in gebessertem Zustand (+). Dies liegt daran, daß im depressiven Zustand des Patienten eher längere Pausen auftreten. Je länger eine zu messende Zeitdauer, desto genauer ist die manuelle Messung. Beispielsweise beträgt die Interraterreliabilität bei der Messung von Äußerungslängen .99 (Pope et al., 1970; Siegman & Pope, 1972).

Abbildung 1 zeigt das Korrelogramm für die Beziehung zwischen den gemeinsamen Pausen.

Dabei wird deutlich, daß auch im Bereich von kürzeren Pausen eine hinreichende Annäherung zwischen manuellen und Computerwerten besteht. Die Homoscedastizität der Verteilung zeigt, daß die hohe Korrelation nicht nur auf der manuellen Messung von längeren Pausen basiert.

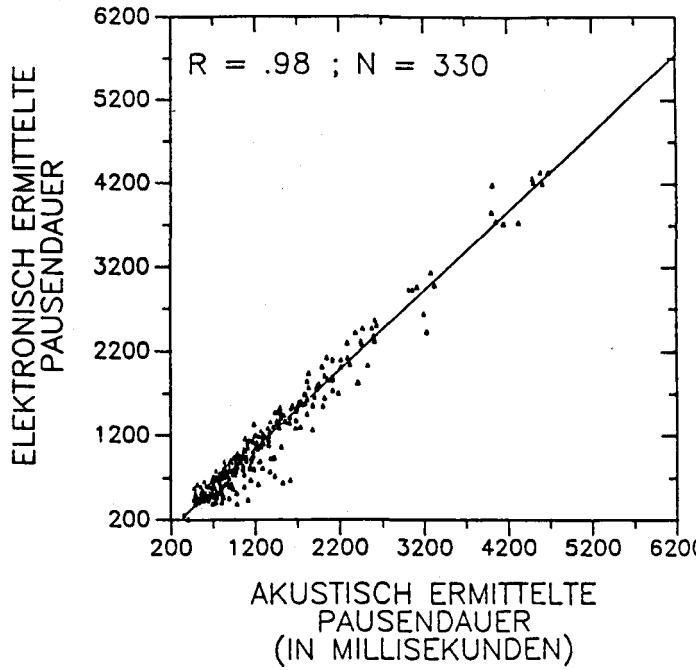

Abb. 1
Übereinstimmung zwischen auditiv und elektronisch gemessenen Stilleperioden.

In Tab. 1 gingen auch die sehr kurzen Stilleperioden zwischen 94 ms und 390 ms aus den Standardfragen der Interviewer ein. Eleminiert man diese und beschränkt sich auf die Sprechpausen oberhalb 390 ms aus der spontanen Sprache von Patienten und Interviewern, ergibt sich folgendes Bild: Von 362 Sprechpausen wurden 304 (84.0%) nach beiden Methoden an gemeinsamer Position im Kontext erfaßt. Zusätzlich brachte die elektronische Auswertung 28 (7.7%), die manuelle Methode 30 (8.3%). Unter Berücksichtigung von Interview 7, welches durch erhöhtes Grundrauschen auch unter Verwendung von verschiedener Filterfrequenzen elektronisch nicht verwertbar war und der 24 vom Rechner durch Amplitudenänderung innerhalb der Stilleperiode fälschlich gespaltenen Pausen, erweist sich die manuelle Methode mit einer Erfassungsrate von 92.6% gegenüber der digitalen Sprachanalyse mit 81.1% aller im untersuchten Korpus enthaltenen Pausen als leicht überlegen. Unter „allen Pausen" verstehen wir die Summe aus G + EZ + AZ. Die „wahre" Zahl der Stilleperioden ist nicht faßbar, da beide Methoden mit einem gewissen Meßfehler behaftet sind.

5. Diskussion

Mit dem vorliegenden Vergleich zwischen Computer und Rater soll nicht etwa die digitale Sprachanalyse in Frage gestellt werden. Vielmehr geht es uns darum zwei Punkte aufzuzeigen: Erstens können Sprechpausen durch eine einfache manuelle Methode ohne technischen Aufwand reliabel gemessen werden. Entgegen den Angaben in der Literatur geht dies auch noch zuverlässig bis zu einem Zeitbereich von 400 ms. Beispielsweise schreiben Siegman & Pope (1972) p. 38: „We decided on two seconds for the lower limit because shorter pauses could not be scored with sufficiently high reliability." Ähnliches behaupten Levin & Silverman (1965), p. 72 für den Bereich unterhalb einer Sekunde. Zweitens sollten Sprachproben zur Vermeidung von Fehlerquellen bei elektronischer Analyse gewisse Bedingungen erfüllen. Bei der Analyse von Sprechpausen mit Methoden der digitalen Sprachverarbeitung sollte das Material frei von Hintergrundrauschen sein. Werden Sequenzen aus Interaktionen untersucht, genügt es nicht mit Krawattenmikrofonen zu arbeiten. Durch Hals- oder Kehlkopfmikrofone sollte gewährleistet sein, daß keine Backchannelsignale des Interaktionspartners im analysierten Signal auftreten. Liegt nur das Sprachsignal von einem Sprecher vor, muß auf Intensitätsschwankungen der Stimme geachtet werden. Leise Endsilben können mit der Amplitude des Hintergrund- bzw. Bandrauschens konkurrieren und zu Störquellen werden, auch wenn sie auditiv noch deutlich erkannt werden. Ebenso sollte die Anzahl von gefüllten Pausen (äh, ehm) eines Sprechers beachtet werden. Sie können wie

Backchannelsignale eines Gesprächspartners Zahl und Dauer der ermittelten Pausen verfälschen.

Bei Sprachproben aus dem klinischen „Feld", z.B. Videoaufnahmen, kann meist nicht Studioqualität auf der Tonspur erreicht werden. Ebenso wirken Hals- oder Kehlkopfmikrofone in vielen Situationen störend auf die natürliche Sprechweise. In diesen Fällen ist die oben beschriebene manuelle Methode empfehlenswert.

Summary

Measuring speech pauses by digital speech analyzing systems yields in poor results under certain conditions. In 16 sandardized interviews with depressive patients speech pauses were analyzed by a simple manual method. The interrater reliability was .88 and greater. These results were compared to those of digital speech analysis. The manual method appeared to be reliable for speech pauses of over 390 ms and reached equal or even better results than digital speech analysis. The manual method is especially profitable in case of soundtracks of bad quality.

Literatur

Clemmer, E. J.: Psycholinguistic aspects of pauses and temporal patterns in schizophrenic speech. Journal of Psycholinguistic Research, 1980, **9**, (2) 161—185.

Ellgring, H., Derbolowsky, J., Dewitz, A. v. & Hieke, S.: Standardisiertes Interview zum Verlauf depressiver Erkrankungen (SID). Max-Planck-Institut für Psychiatrie—Sozialpsychologie—, München 1978.

Greden, J. F. & Carroll, B. J.: Decrease in speech pause times with treatment of endogenous depression. Biological Psychiatry, 1980, **15**, 575—587.

Greden, J. F., Albala, A. A., Smokler, I. A., Gardner, R. & Carroll, B. J.: Speech pause time: a marker of psychomotor retardation among endogenous depression. Biological Pychiatry, 1981, **16**, 851—859.

Helfrich, H.: A digital method of pause extraction. In: H. W. Dechert & M. Raupach (Eds.) Temporal variables in speech. Mouton, N. Y. 1980.

Klos, K.-T. & Ellgring, H.: Sprechgeschwindigkeit und Sprechpausen von Depressiven. In: M. Hautzinger & R. Straub (Hrsg.) Psychologische Aspekte depressiver Störungen. Regensburg: S. Roderer.

Kuenzel, H.: Eine experimentelle Untersuchung zur Pausenperzeption. Arbeitsberichte des Instituts für Phonetik, Kiel (AIPUK) 1974, **2**, 1—64.

Levin, H. & Silverman, I.: Hesitation phenomena in children's speech. Language and Speech, 1965, **8**, 67—85.

Mawdsley, C. & Gamsu, C. V.: Periodicity of speech in parkinsonism. Nature, 1971, **231**, 315—316.

Pope, B., Blass, T., Siegman, A. W. & Raher, J.: Anxienty and depression in speech, 1970. Journal of Consulting and Clinical Psychology, **35**, 128—133.

Scherer, K. R.: Methods of research on vocal communication: Paradigms and parameters. In: K. R. Scherer & P. Ekman (Eds.) Handbook of methods in nonverbal behavior research. Cambridge University Press, 1982, 136—198.

Siegman, A. W. & Pope, B.: The effects of ambiguity and anxiety on interviewee verbal behavior. In: A. W. Siegman & B. Pope (Eds.) Studies in dyadic communication. N. Y., Pergamon Press, Kap. 3, 29—67.

Standke, R.: Gisys: Ein Software-Editor zur fileorientierten digitalen Sprachverarbeitung im Zeitbereich. In: W. Michaelis (Hrsg.) Bericht über den 32. Kongreß der Deutschen Gesellschaft für Psychologie in Zürich. Bd. I. Hogrefe, 1980, Göttingen 197—200.

Anschriften der Autoren: Dipl.-Psych. K.-T. Klos, Zentralinstitut für Seelische Gesundheit, Postfach 5970, 6800 Mannheim 1, und PD Dr. H. Ellgring, Max-Planck-Institut für Psychiatrie, Kraepelinstraße 10, 8000 München 40.