Phonetische Beiträge zur maschinellen Spracherkennung

DIE EXTREMWERTANALYSE IM ZEITBEREICH ALS MÖGLICHKEIT DER AUTOMATISCHEN SPRACHERKENNUNG

Thomas und Patrick Schweisthal [1]

Arbeitskreis Spracherkennung, Sprachgenerierung und Phonetische Datenbanken
der Gesellschaft für Linguistische datenverarbeitung (GLDV)
Institut für Phonetik und sprachliche Kommunikation
Universität München
Schellingstraße 3 V6
8000 München 40

Zukünftige Computersysteme sollten in der Lage sein, gesprochene Sprache ohne die bisher üblichen Einschränkungen (wie lange Trainingsphasen für Sprecher bei geringem Wortschatz) zu erkennen. Der Beitrag des Phonetikers hierzu kann eine Transkiptionsmaschine sein, die gesprochene Sprache in phonetische Zeichen umsetzt. Im nachfolgenden Artikel wird unter Berücksichtigung neuerer Erkenntnisse aus der Physiologie des Gehörs die Extremwertanalyse im Zeitbereich als Möglichkeit der akustischen Sprachschallanalyse dargestellt. Inwiefern sie eine Alternative zu gängigen Verfahren darstellt, wird sich bei genauerer Erprobung zeigen.

Seit langer Zeit bemüht sich die Forschung um ein zufriedenstellendes Verfahren zur Automatischen Spracherkennung. Für isolierte Worterkennung ist man diesem Ziel schon sehr nahe gekommen, während die Untersuchung fortlaufender Rede bisher unter unverhältnismäßigem Aufwand (Aktas/Kämmerer/Küpper/Lagger 1986 S. 9 *"Trotz vieler neuer Ansätze auf dem Gebiet der Spracherkennung (Verwendung von Markov-Modellen, auf phonetischem Wissen basierende Systeme) ist der klassische Mustererkennungssatz mit dynamischer Programmierung immer noch von vorrangiger Bedeutung... Bei großen Wortschätzen (mehr als 1000 Wörter) ist es jedoch schwierig, mit diesen sehr rechenintensiven Verfahren Echtzeitbedingungen einzuhalten."*) noch unbefriedigende Ergebnisse vorzuweisen hat.

Wir glauben, daß man mit den bisherigen Mitteln der Spektralanalyse, in die die größten Hoffnungen gesetzt worden sind und noch werden, noch nicht den richtigen Weg einer akustischen Sprachanalyse eingeschlagen hat; also werden auch andere Bereiche des Signalphonetischen Bandes (*Tillmann 1980, S. 193*) zur Spracherkennung herangezogen, obwohl nach unserer Auffassung das akustische Signal die notwendige Information für eine am Hören orientierte Sprachschallanalyse enthält (Pompino 1978, S. 22 f. *"Meines Erachtens ist der Produktionsmechanismus nicht die Grundlage der Perzeption, der Perzeptionsmechanismus vielmehr ist auditorisch schon so ausgestattet, daß sich der Zusammenhang zwischen Produktion und Perzeption ontogenetisch formen kann."*).

Die Spektralanalyse zerlegt das Zeitsignal in seine Bestandteile verschiedener Frequenzbereiche, in denen Energiekonzentrationen festzustellen sind. Daß aber das menschliche Ohr nicht nur wie die Filterbank eines Sonagraphen funktioniert, ist längst nachgewiesen; lediglich für Frequenzanteile über 5000 Hz mag das in etwa zutreffen. Für den Frequenzbereich unter 500 Hz wird eine direkte zeitliche Wahrnehmung angenommen, für den verbleibenden Bereich von etwa 500 bis 5000 Hz eine gemischte Zeit-Ort- Wahrnehmung (Volley-Theory) (vgl. *Schouten 1938, 1940; Ainsworth 1976, S. 31 ff; Spreng 1975, S. 138*). Dem akustischen Rezeptionsapparat Ohr also wird die Spektralanalyse allein nicht gerecht, zumal bekannte Physiologen der Sprachforschung den Rat gegeben haben, sich verstärkt mit der Zeitanalyse zu befassen (Spreng 1975, S. 138 *"Alle diese Befunde zeigen, daß neben der reinen Frequenzerkennung durch entsprechende Ortsabbildung eine Zeitanalyse vor allem bei der Erkennung der Sprache eine entscheidende Rolle spielt, so daß man mit Recht als die hervorstechendste Eigenschaft des Gehörs die Analyse zeitlicher Muster bezeichnen kann."*).

Schon im Jahre 1951 ermittelte man Nulldurchgangsdichten des Originalsignals und des differenzierten Signals (also die Extremwertdichte des Ausgangssignals), *(Chang / Pihl / Essigmann 1951)*, nachdem zuvor Licklieder und Pollak *(Licklider / Pollak 1948; Ainsworth 1976, S. 97 f)* in Untersuchungen mit manipulierten Sprachsignalen darauf hingewiesen hatten, daß ein Großteil sprachlicher Information in den Extremwertabständen liegen müsse.

Peterson stellte des weiteren die bemerkenswerte Proportionalität von Nulldurchgangsdichte zum ersten Formanten und Extremwertdichte zum zweiten Formanten der Spektralanalyse heraus *(Peterson 1951)*. Es bestand aber weithin Einigkeit, daß diese Verfahren im Zeitbereich zu eher schlechteren Ergebnissen führten als die Spektralanalyse *(Kirstein 1977, S. 11)*. Eine Untersuchung zum Vergleich zwischen Formantextraktion durch Filterbänke und andererseits durch Nullstellenanalyse lieferte *Scarr 1968*, wenn auch spätere Untersuchungen dieser Ansicht widersprechen *(Kirstein 1977, S. 4 ferner S. 99)*. Kirstein, der selber nach Nullstellen analysiert, empfiehlt, Extremwertabstände in jedem Fall in weitere Untersuchungen einzubeziehen – weiteres dazu siehe Literaturverzeichnis.

Unter Berücksichtigung dieser Erkenntnisse bieten unserer Ansicht nach die Extremwerte den besten Ansatzpunkt für die Zeitanalyse: Im Gegensatz zur Nullstellenanalyse verändert sich die Extremwertlage nicht mit einem veränderten Berechnungsverfahren der Nullinie, die ja den über einen bestimmten Zeitraum gemittelten Wert wiedergibt.

Fraglich ist allerdings, wie man dem akustischen Apparat Ohr bei der Analyse am ehesten gerecht werden kann. Erhalten bleiben sollen bei einem solchen Verfahren jedenfalls die ursprünglichen Extremwertabstände, was bei herkömmlichen (linearen) Hoch- und Tiefpässen nicht der Fall ist, wohl aber bei einem System, das die Kurve in Grob- und Fein-

strukturen aufteilt, wobei die Grobstruktur den Bereich der rein zeitlichen Wahrnehmung (bis etwa 500 Hz), die Feinstruktur den der Volley-Theory (etwa von 500 bis 5000 Hz) vertritt. Das läßt sich aber über ein herkömmliches Glättungsverfahren erreichen, welches die einzelnen Punkte der digitalisierten Kurve neu definiert als die Mittelwerte ihrer Umgebungspunkte der Originalkurve, ein nichtlineares Filterverfahren also. Es ergibt sich eine neue Kurve, für die in einem festgesetzten zeitlichen Bereich die Gesamtdifferenz der neuen Werte zu den Originalwerten gleich null ist, wobei die neuen Werte möglichst wenig von ihren Nachbarwerten abweichen. Die Differenz aus der so gewonnenen Grobstruktur und der Originalkurve ergibt die Feinstruktur. Wie weit der Mitteilungszeitraum am sinnvollsten zu setzen ist, ist experimentell zu bestimmen. Bei Grob- und Feinstruktur ist nebeneinander der zeitliche Abstand benachbarter Extremwerte festzuhalten. Die Reziprokwerte der Abstände, die "Extremwertfrequenzen", in Abhängigkeit von der Signalzeit abgetragen, ergeben ein Sonagrammen nicht unähnliches Bild, weshalb wir Häufungen bestimmter Extremwertfrequenzen als die wirklichen Formanten, die "Realformanten" bezeichnen möchten. Ein vergleichbares Verfahren wandte Janet M. Baker *(Baker 1974)* an.

Wir konnten, natürlich unter Vernachlässigung mancher Unregelmäßigkeiten - entsprechend der Formantanalyse im Spektralbereich - folgende qualitative Zuordnungen der Zungenlage und Zungenhöhe formulieren:

- je tiefer die Zunge, desto höher die Extremwertfrequenz der Grobstruktur (des ersten Realformanten)
- je weiter vorn die Zunge und je höher die Zunge, desto höher die Extremwertfrequenz der Feinstruktur (des zweiten und eventuell weiterer Realformanten)

Für die gesamte hintere Reihe (u bis hinteres tiefes a) bedeutet das, daß sie durch den ersten Realformanten hinreichend charakterisiert ist.

Für den Zentralbereich gilt, daß sich die Werte des ersten und zweiten Realformanten einander annähern. Auf dieser Basis wollen wir eine "Transkriptionsmaschine" bauen.

Unsere Bemühungen werden sich dahin richten, jedem möglichen Wertepaar der Realformanten 1 und 2 ein Schriftzeichen zuzuordnen und hierzu eindeutige, lückenlose Definitionsbereiche zu finden. Jedes Segment ist somit durch ein Wertepaar definiert (Objektive Transkription). Im zweiten Arbeitsgang wird unter Berücksichtigung von Koartikulationserscheinungen, artikulorischer Nachbarschaft und psychoakustischen Phänomenen eine "Subjektive Transkription" versucht, die der üblichen Unterrichtstranskription entsprechen soll. (Beispiel für eine objektive Transkription eines /ta/:

[/i/e/ä/a/.../a/];

die ersten drei Zeichen /i/e/ä/ stellen eine charakteristische Transition von /t/ zu /a/ dar, lassen sich also subjektiv als t vor a transkribieren. Die objektive Transkription transkribiert im stimmhaften Bereich periodenweise, im stimmlosen Bereich segmentweise.

Ist erst die "Hürde" einer brauchbaren unkomplizierten Transkriptionsmaschine genommen, so wird unter Berücksichtigung weiterer sprachlich relevanter Gegebenheiten wie Tonhöhe und Intensitätsbetrachtung über längere Zeiträume eine weitere Unterteilung auf Silbenebene und in der Endstufe durch weitergehende Anwendung von Methoden der Künstlichen Intelligenz eine Verschriftung schließlich möglich sein. Für die Spracherkennung ergeben sich folgende Einzelschritte:

1. Digitalisierung des Sprachsignals
2. Objektive Transkription
3. Subjektive Transkription (Verschriftung nac Lautschrift)
4. Versilbung (unter Berücksichtigung längere Intensitätsverläufe und der Tonhöhe)
5. Syntaktische Analyse und schließlich
6. Verschriftung

Anmerkung

[1] Die Autoren sind auch unter folgender Privatadresse z erreichen: Thomas und Patrick Schweisthal, Edenthalweg 3 8069 Rohrbach, Tel. 08442-8444.

Literatur

Ainsworth, W.A. 1976: Mechanisms of Speech Recogni tion, 1976

Aktas, A. / Kämmerer, B. / Küpper, W. / Lagger, H 1986: Spracherkennung für große Wortschätze mit schneller grober Zeitanpassung. In NTG- Fachberichte 94, 1986

Baker, J.M. 1974: A new time domain analysis of fricatives and stop consonants. In Proc. IEEE Symp. Speech Recognition, Pittsburgh 1974, S. 134-141

Chang, S.H. / Phil, G.E. / Essigmann, M.W. 1951 Representation of Speech Sounds and some of their statistica properties. In Proc. Inst. Radio Engrs. 39, 1951, S. 147-152

Ito, M.R. 1972: Relationship between zero-crossing measurements for speech analysis and recognition. J. Acoust. Soc Am. 51, 1972, S. 2061-2062

Keller, Th., von 1976: Die Kennzeichnung von Sprachlauten durch Spektrum, Autokorrelationsfunktion und Nulldurchgangsabstände. Nachrichtentechn. Z. 20, 1976, S. 195-205

Keidel, W.D. 1970: Kurzgefaßtes Lehrbuch der Physiologie. Thieme, Stuttgart, 1970 (3. Aufl. 1973)

Kirstein, M. 1977: Untersuchung zur Verteilung von Nulldurchgangsabständen bei Sprachsignalen, Buske, Hamburg 1977

Kotten, K. 1971: Nulldurchgangszählung und Nulldurchgangsdichte in DAWID-II. Beiträge zur automatischer Spracherkennung. Forschungsberichte IPK, Band 36, Hamburg 1971, S. 81-111

Licklider, J.C.R. / Pollak, I. 1948: Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech. J. Acoust. Soc. Am. 20, 1948, S. 42-51

Meyer-Eppler, W. 1969: Grundlagen und Anwendungen der Informationstheorie. Zweite Auflage Berlin, 1969

Peterson, E. 1951: Frequency detection and speech formants. J. Acoust. Soc. Am. 23, 1951, S. 668, 674

Peterson, G.E. / Hanne, J.R. 1965: Examination of two different formant estimation techniques. J. Acoust. Soc. Am 38, 1965, S. 224-228

Pompino, B. 1978: Hemisphärenunterschiede und der 'phonetic speech processor'. In Forschungsberichte 9, 1978 Institut für Phonetik und Sprachl. Kommunikation, München

Reddy, D.R. 1967: Computer recognition of connected speech. J. Acoust. Soc. Am. 42, 1967, S. 329-347

Scarr, W.A. 1968: Zero crossings as ameang of obtaining spectral information in speech analysis. IEEE Trans. Audio Electroacoust. Au-16, 1968, S. 247-255

Schouten, J.F. 1940: The perception of pitch. Philips techn. Rev. 5, 1940, S. 286-294

Spreng, M. 1975: Mechanisch-elektrische Wandlung: Dynamik der Sinneszellen des Gehörs, Adaption und selektives Verhalten. In: Keidel W.D.(Hrsg): Physiologie des Gehörs Stuttgart 1975, S. 102-152