Phonetische Beiträge zur maschinellen Spracherkennung

Parameter der Sprachanalyse im Zeitbereich ein Schritt in Richtung Transkriptionsmaschine?

Patrick Schweisthal Walter Kopetzky Arbeitskreis Spracherkennung, Sprachgenerierung und phonetische Datenbanken der Gesellschaft für linguistische Datenverarbeitung Institut für Phonetik und sprachliche Kommunikation Universität München, Schellingstr. D-8000 München 40

In den beiden letzten Ausgaben des LDVForum wurde die Extremwertanalyse im Zeitbereich als Verfahren der automatischen Sprachanalyse vorgestellt; anhand ausgewählter Intervallogramme und Sonagramme derselben Äußerungen erfolgte eine Gegenüberstellung von Zeitbereichs- und Frequenzbereichsanalyse. Der vorliegende Beitrag behandelt, ein von den Autoren entwickeltes Zeitanalyseverfahren. Anhand des Zeitsignals wurden Parameter herausgearbeitet, die eine Zuordnung von akustischem Produkt und Wahrnehmungskategorien ermöglichen sollen. Besonders zu beachten ist der Stimmhaftigkeits- und Tonhöhenalgorithmus, der in leicht abgewandelter Form auch zur Beschreibung der Realformanten herangezogen wird.

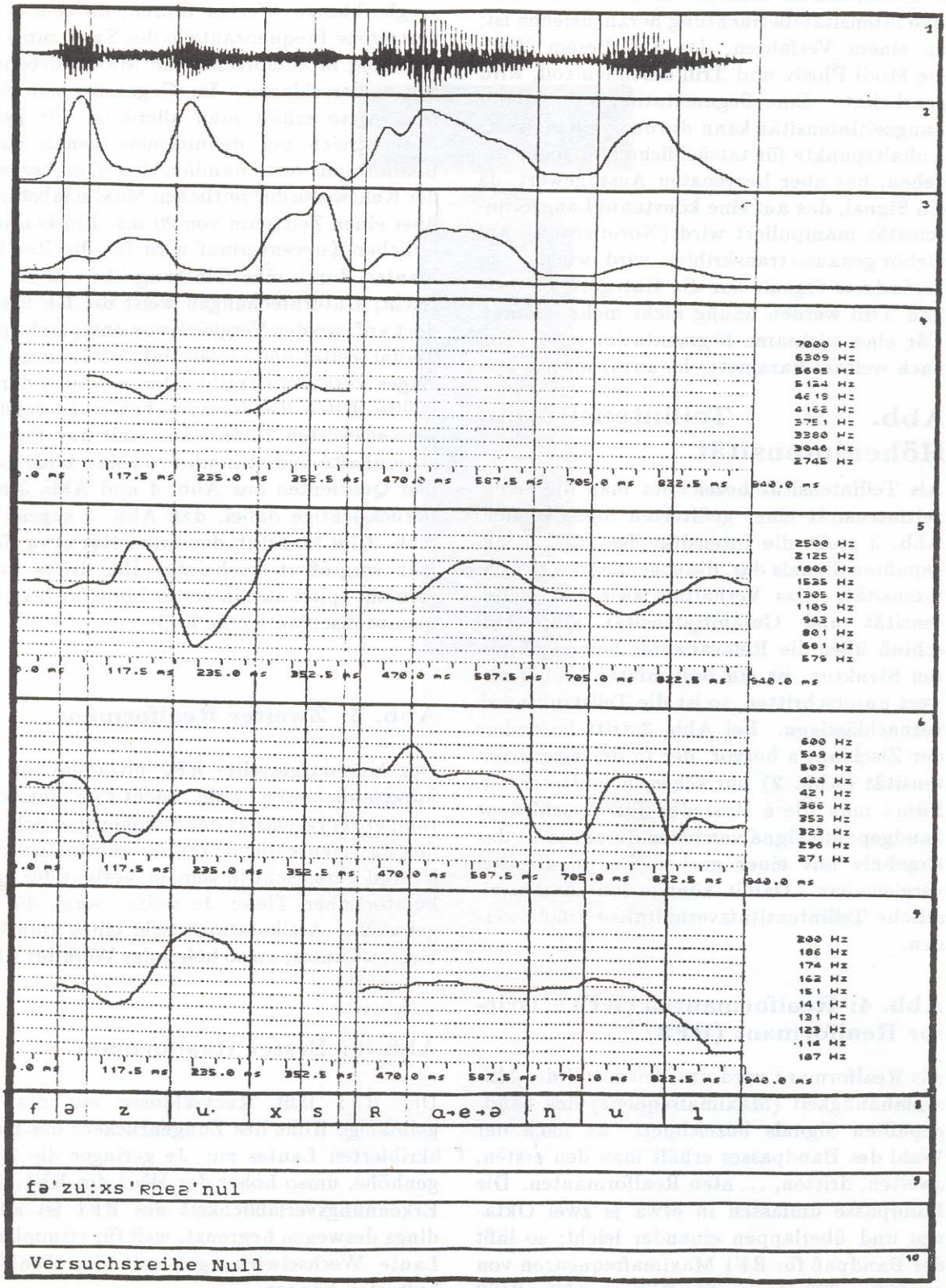
Die Parameter sind in zwei Gruppen zu unterteilen, nämlich Intensitätsbetrachtungen (Abb. 2 und 3) und Extremwertdichtebetrachtungen (Abb. 4 bis 7). Im folgenden werden die Parameter erklärt, wobei auf ihre mögliche Bedeutung für die Spracherkennung hingewiesen wird.

Abb. 1: Zeitsignal, Oszillogramm

Das Zeitsignal stellt die Druckveränderung in Abhängigkeit von der Zeit dar, die wir als Schall wahrnehmen. Die Form des Zeitsignals kann für dieselbe Äußerung - abhängig von Aufnahmedecoder und Mikrofoneigenschaften - erheblich variieren; wegen dieser hohen Varianz hat das Zeitsignal den Ruch, für die Sprachanalyse weitgehend ungeeignet zu sein, zumal entsprechende Verfahren, die sich auf den gesamten Kurvenverlauf beziehen, den Eindruck der hohen Varianz noch bekräftigen und einen unpraktikablen Rechenaufwand mit sich bringen. Demgegenüber sind geübte Transkribenten, die sich regelmäßig mit Computersegmentation an Sprachsignal beschäftigen, mit Einschränkungen in der Lage, das Oszillogramm ohne akustische Wiedergabe zu "lesen". Das brachte die Autoren auf die Idee, sich den Signalpartien zuzuwenden, an denen sich der Segmentator orientiert, nämlich den Extremwerten. Die' vorliegende Darstellung ist allerdings zeitlich so gestaucht, daß Einzelheiten nicht erkennbar sind. Der maschinellen Sprachanalyse wird das Zeitsignal durch die Auf teilung in Unterstrukturen (Realformantbereiche), Intensitäts- und Extremwertbetrachten zugänglich.

Abb. 2: Langzeitintensität, Gesamtintensität

Die Langzeitintensität stellt ein Zeitmittel der im Oszillogramm angezeigten Druckveränderungen dar. Das hier verwendete Verfahren richtet das Oszillogramm gleich und mittelt die so erhaltene Kurve zweimal über einen Zeitraum von 20 ms. Zu beachten ist, daß die absolute Intensität für die Spracherkennung nur insofern Bedeutung hat, als bestimmte Grenzwerte nach oben (Lärm) und nach unten (Stille) nicht überschritten werden können, ohne die Wahrnehmung zu beeinträchtigen; die absolute Intensität hängt schließlich von


1
2
3
4
7000 Hz
6309 Hz
5685 Hz
5124 Hz
4619 Hz
4162 Hz
3751 Hz
3380 Hz
3046 Hz
2745 Hz
5
2500 Hz
2125 Hz
1806 Hz
1535 Hz
1305 Hz
1109 Hz
943 Hz
801 Hz
681 Hz
579 Hz
6
600 Hz
549 Hz
503 Hz
460 Hz
421 Hz
386 Hz
353 Hz
323 Hz
296 Hz
271 Hz
7
200 Hz
186 Hz
174 Hz
162 Hz
151 Hz
141 Hz
131 Hz
123 Hz
114 Hz
107 Hz
0.0 ms
117.5 ms
235.0 ms
352.5 ms
470.0 ms
587.5 ms
705.0 ms
822.5 ms
940.0 ms
8
f ə z u· x s ʁ a→e→ə n u l
9
fə'zu:xs'ʁaeə'nul
10
Versuchsreihe Null

den Aufnahmebedingungen ab; auch ist die Intensitätswahrnehmung hörer- und frequenzabhängig. Aus diesen Gründen ist ersichtlich, daß als Maßstab in erster Linie relative Intensitätsbetrachtung heranzuziehen ist; an einem Verfahren, das auf diesem Wege die Modi Plosiv und Trill erkennen soll, wird gearbeitet. Eine Segmentation anhand der Langzeitintensität kann darüber hinaus zwar Anhaltspunkte für tatsächliche Lautsegmente geben, hat aber begrenzten Aussagewert, da ein Signal, das auf eine konstante Langzeitintensität manipuliert wird (Normierung), auf Gehör genauso transkribiert wird, wie das unveränderte Signal; nur die Kategorien Plosiv und Trill werden häufig nicht mehr erkannt. Für eine wirksame Segmentation sind demnach weitere Parameter heranzuziehen.

Abb. 3: Teilintensitäten, Höhenintensität

Als Teilintensität bezeichnet man die Langzeitintensität einer gefilterten Signalversion. Abb. 3 stellt die Intensität des stark hochgepaßten Signals dar, die sogenannten Höhenintensität. Das Verhältnis zwischen Teilintensität und Gesamtintensität gibt Aufschluß über die Relevanz der herausgefilterten Struktur; ist ein bestimmter Verhältniswert unterschritten, so ist die Teilstruktur zu vernachlässigen. Bei Abb. 3 tritt besonders der Zischlaut s hervor, der in der Gesamtintensität (Abb. 2) nur schwach vertreten ist. Führt man diese Prozedur für verschiedene bandgepaßte Signalversionen durch, so ist das Ergebnis mit einer groben Fourier-Section vergleichbar. Damit können lautcharakteristische Teilintensitätsverhältnisse erfaßt werden.

Abb. 4: Realformanten (RF) - Dritter Realformant (RF3)

Als Realformant wird das Zeitmittel der Maximahäufigkeit (Maximafrequenz) des bandgepaßten Signals bezeichnet. Je nach der Wahl des Bandpasses erhält man den ersten, zweiten, dritten, .. . nten Realformanten. Die Bandpässe umfassen in etwa je zwei Oktaven und überlappen einander leicht; so läßt der Bandpaß für RFl Maximafrequenzen von ca. 200 - 800 Hz zu, für RF2 ca. 600 - 2400 Hz, für RF3 ca. 2000 - 8000 Hz. Zum verwendeten Filter siehe LDV-Forum Juni 87. Im Vergleich mit den Fourierformanten ist festzustellen, daß die Realformantverläufe zu vergleichbaren Werten führen, da sich stark vertretene Frequenzanteile des Spektrums regelmäßig in entsprechenden Maximafrequenzen niederschlagen. Im Gegensatz zur Fourieranalyse erhält man allerdings für jeden Filterbereich per definitionem genau einen Realformantwert, nämlich den Durchschnitt der Kehrwerte der zeitlichen Maximaabstände über einen Zeitraum von 20 ms. Ein kontinuierlicher Kurvenverlauf wird für die Realformanten durch eine Glättung über 20 ms erreicht; Unterbrechungen weist die RF-Kurve dort auf, wo der Vergleich von entsprechender Teilintensität und Gesamtintensität ein zu geringes Verhältnis ergibt. Der in Abb. 4 dargestellte dritte Realformant hat einen solchen nennenswerten Intensitätsanteil nur für das stimmhafte und stimmlose s (Man bilde dazu den Quotienten aus Abb. 4 und Abb. 3 und berücksichtige dabei, daß Abb. 4 gegenüber Abb. 3 im Maßstab der Intensität etwa fünffach vergrößert erscheint). Der dritte Realformant spielt also in erster Linie für hochfrequente Zischlaute eine Rolle.

Abb. 5: Zweiter Realformant

Der hier dargestellte RF2 birgt die für die Spracherkennung wohl wichtigste Information; er repräsentiert den Frequenzbereich, für den das menschliche Ohr am empfindlichsten ist und entspricht in seinem Verlauf der artikulatorischen Tiefe: Je weiter vorne die regelmäßige Artikulationsstelle eines transkribierten Lautes, umso höher der Wert des RF2.

Abb. 6: Erster Realformant

Der RFl läßt Rückschlüsse auf die regelmäßige Höhe des Zungenrückens des transkribierten Lautes zu: Je geringer die Zungenhöhe, umso höher der Wert des RFl. Die Erkennungsverläßlichkeit des RFl ist allerdings deswegen begrenzt, weil für stimmhafte Laute Wechselwirkungen mit der Tonhöhe auftreten können.

Abb. 7: Die Tonhöhenkurve

Die Tonhöhenkurve gibt das Zeitmittel der Kehrwerte der zeitlichen Dauern quasiperiodischer Einheiten der Signalkurve wieder; meßbar werden die Periodendauern durch einseitige Gleichrichtung des Oszillogramms und eine Kombination spezieller Digitalfilter gemessen werden die Maximaabstände der so gewonnen Kurve. Anschließend entscheidet ein Auswahlalgorithmus über die Periodizität: N ur die Abstände werden berücksichtigt, die den Periodizitätsanforderungen gerecht werden; aus ihren Kehrwerten wird eine Kurve erstellt, die zweimal über 20 ms geglättet wird. Das Ergebnis entspricht einer Periodensegmentation von Hand. Das Tonhöheverfahren arbeitet für einen Bereich von bis zu zwei Oktaven sehr zuverlässig und liefert damit verbindliche Angaben über Stimmhaftigkeit und Betonung.

Abb. 8: Segmentelle Transkription

Hier wurden Segmenten nach Gehör des Segmentators Lautqualitäten zugeordnet; Ziel der Arbeitsgruppe ist es, eine solche Transkription vom Rechner erstellen zu lassen.

Abb. 9: Phonetische Transkription

Abb. 9 zeigt die fertige phonetische Transkription mit Akzenten als adäquate Verschriftung der Äußerung.

Abb. 10: Orthographische Umschriftung

Das Ziel der Spracherkennung ist, fließend gesprochene Sprache automatisch in ihre orthographische Norm zu überführen. Das kann unserer Ansicht nach aber nur über eine praktikable, automatische, phonetische Transkription gelingen.