Automatische Segmentierung der Lautelemente

Dr.-Ing. Shahla Sehhati, TU-Berlin

shahla@prz.tu-berlin.de

Kurzfassung

Das Programmpaket zur automatischen Segmentierung bestimmt auf Basis von Spektralvergleichen die Lautgrenzen. Das Verfahren ist für alle Lautelemente wie Phoneme, Diphone, Cluster etc. universell einsetzbar.

Es werden in Abhängigkeit der verwendeten Referenzspektren die verschiedenen Laute extrahiert. Zur Illustration des Verfahrens werden Cluster verwendet.

Das Programmpaket verwendet bei der Segmentierung implizite und explizite Verfahren und führt anschließend eine Kombination der beiden aus.

Bei der Kombination der beiden Methoden werden nur die Lautgrenzen übernommen, die durch beide Verfahren ermittelt worden sind.

Die nach diesem Verfahren gefundenen Lautgrenzen liefern bis zu 95 % genaue Ergebnisse.

Automatisch segmentierte Signale können grafisch dargestellt und nachträglich manipuliert werden.

Arbeitsweise der Verfahren

Implizites Verfahren

Dieses Verfahren arbeitet in drei Schritten. Zunächst wird mit Hilfe der FFT eine Folge von Kurzzeitspektren berechnet und die Ähnlichkeitsbeziehung zwischen benachbarten Kurzzeitspektren hergestellt, um die Lautübergänge zu bestimmen. Die Korrelation läßt sich mit folgender Gleichung bestimmen:

$$C_{ij} = \frac{(S_i \cdot S_j)}{\sqrt{(S_i \cdot S_i)\cdot(S_j \cdot S_j)}} \qquad \text{(Gl. 1)}$$

Es läßt sich das Skalarprodukt von S_i und S_j für die stetige Funktion definieren:

$$(S_i \cdot S_j) = \int_0^{8KHz} W(f)\cdot S_i(f)\cdot S_j(f)\, df \qquad \text{(Gl. 2)}$$

mit der Gewichtungsfunktion

$$W(f) = \begin{cases} 0 & 0 < f \le 200\ Hz \\ 1 & 200 < f \le 1000\ Hz \\ 1000/f & f > 1000\ Hz \end{cases} \qquad \text{(Gl. 3)}$$

entsprechend ergibt sich die normierte Gleichung:

$$\| S_i \| = \sqrt{(S_i \cdot S_i)} \qquad \text{(Gl. 4)}$$

Es gilt dabei die Schwarzsche Ungleichung:

$$\left|(S_i \cdot S_j)\right| <= \| S_i \| \cdot \| S_j \| \qquad \text{(Gl. 5)}$$

Es werden im *ersten* Schritt zu jedem Kurzzeitspektrum S*i* alle benachbarten Kurzzeitspektren S*j* ermittelt, die S*i* ähnlich sind. Es wird ein Schwellwert C als Vergleichswert für die Ähnlichkeit herangezogen; seine Größe kann experimentell ermittelt werden (Bild 1).

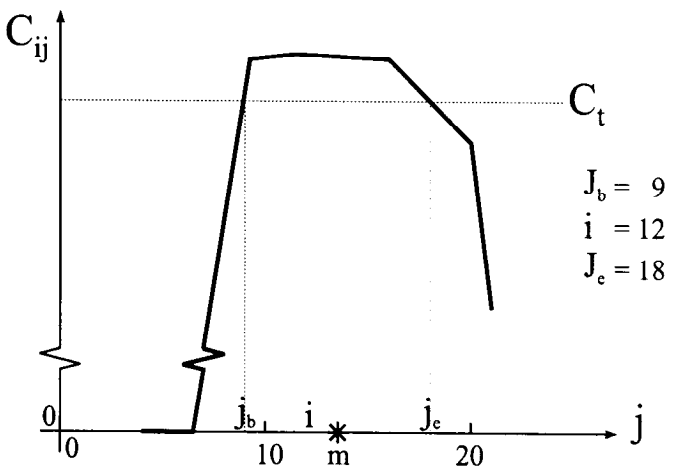

Bild 1: Die berechneten Korrelationen C_{ij} zum Kurzzeitspektrum *i* (=12) und seinen ähnlichen Nachbarspektren; C_t ist der Schwellwert, j_b : der Segmentbeginn und j_e : das Segmentende

Liegt der berechnete Korrelationswert unter dem Schwellwert, so sind die betrachteten Spektren nicht ähnlich; liegt er darüber, werden sie als ähnlich angesehen. Alle zu einem Kurzzeitspektrum S_i ähnlichen Nachbarspektren S_j werden zu einem Segment zusammengefaßt.

Im *zweiten* Schritt werden Schwerpunkte m_i eines jeden Segments berechnet:

$$m_i = \frac{\sum_{j=i_b}^{i_e} j(C_{i,j} - C_t)}{\sum_{j=i_b}^{i_e} (C_{i,j} - C_t)} \qquad \text{(Gl. 6)}$$

In einem *dritten* Schritt wird die Entfernung *d* des Kurzzeitspektrums *i* zum Schwerpunkt *m* wie folgt berechnet:

$$d_i = m_i - i \qquad \text{(Gl. 7)}$$

Die berechnete Differenz *d* zeigt bei geeigneter Wahl des Schwellwertes *C* einen Wechsel der Werte vom positiven in den negativen Bereich und umgekehrt.

Der Nulldurchgang vom positiven in den negativen Bereich kennzeichnet ein Kurzzeitspektrum, das sich im Schwerpunkt eines Segmentes befindet. Der Nulldurchgang vom negativen in den positiven Bereich weist auf ein Wechsel des quasistationären Zustands hin und markiert somit eine Lautgrenze.

Explizites Verfahren

Dieses Verfahren verwendet Referenzspektren zur Analyse des Sprachsignals. Damit ein im Sprachsignal enthaltener Laut mit Sicherheit entdeckt wird, führt man Einschränkungen für die Länge eines Spektralzustandes ein. Es wird ermöglicht, daß ein

vorgesehener Zustand bei spektraler Unähnlichkeit mit dem momentan abgeprüften Zustand auf seine minimale Länge segmentiert wird.

Die Einführung einer maximalen Länge führt dazu, daß bei großer spektraler Ähnlichkeit nach Ablauf der maximalen Zeitdauer zwangsläufig der nächste vorgegebene Laut segmentiert wird. Zur Ermittlung dieser Minima und Maxima wurde die Länge von Clustern eines Sprechers zugrundegelegt. Die Abweichung der Clusterlänge beträgt ca. 10 % ; sie wurde bei der Erstellung der Referenzbibliothek berücksichtigt.

Es wird eine Matrix der Größe $i \cdot n$ erzeugt, wobei i die Anzahl der Referenzspektren und n die Anzahl der Kurzzeitspektren des zu analysierenden Sprachsignals darstellt. Die Korrelation jedes im Zeitsignal enthaltenen Referenzspektrums mit allen Kurzzeitspektren wird berechnet.

Die eigentliche Segmentierung besteht nun darin, die Grenzen durch das Maximieren der Korrelation C zwischen dem i-ten Kurzzeitspektrum und dem n-ten Referenzspektrum zu finden. Für das Maximum gilt:

$$\max \sum_{n=1}^{N} \sum_{i=g_{n-1}+1}^{g_n} C_{i,n} \qquad \text{(Gl. 8)}$$

mit der Bedingung

$$\min_n \le g_n - g_{n-1} \le \max_n \quad \text{für} \quad n = 1 \ldots N, \qquad \text{(Gl. 9)}$$

wobei

n	der Index des Referenzspektrums,
N	die Anzahl der Referenzspektren,
i	die Nummer des Kurzzeitspektrums,
I	die gesamte Anzahl der Kurzzeitspektren,
g_n	das letzte Kurzzeitspektrum des n-ten Referenzspektrums und
$\min_n/\max_n$	die minimale und die maximale Dauer des n-ten Lautes sind.

Kombination beider Verfahren

Durch Kombination dieser Methoden werden Lautgrenzen ermittelt (Bild 2).

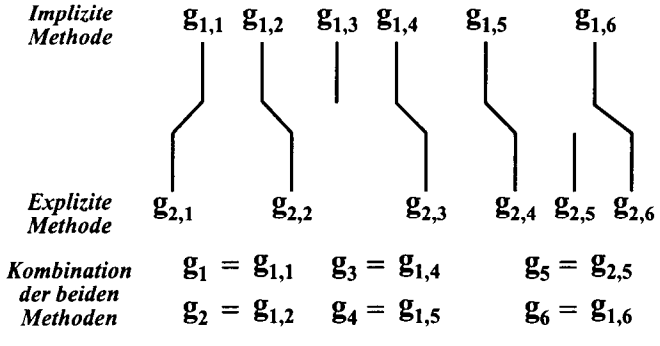

Bild 2: Kombination der beiden Methoden

Zwei Lautgrenzen werden verbunden, wenn sie in der Nähe liegen. Die Lautgrenze nach dem ersten Verfahren, zu der sich keine Grenze mit dem zweiten Verfahrens zuordnen läßt, wird ausgelassen. Die Lautgrenzen nach der zweiten Methode, zu denen sich nach der ersten Methode keine Grenzen finden lassen, bleiben als endgültige Grenzen erhalten. Durch diesen Schritt wird die Anzahl der Laute der ersten Methode und die Ungenauigkeit der gefundenen Grenzen der expliziten Segmentierung korrigiert.

Zusammenfassung

Im Rahmen dieser Arbeit ist ein Programmteil zum Erstellen und Verwalten der unterschiedlichsten Referenzspektren realisiert worden.

Die nach diesem Verfahren gefundenen Lautgrenzen liefern bis zu 95 % genaue Ergebnisse.

Automatisch segmentierte Signale können grafisch dargestellt und nachträglich manipuliert werden. Das Bild 3 zeigt richtig gefundene Lautgrenzen bei einem Schwellwert von 0.98.

Zur Korrektur der gefundenen Lautgrenzen und zur Verwaltung und Analyse der segmentierten Sprachdaten arbeitet das Programmpaket mit dem Programmpaket SPANEX zusammen.

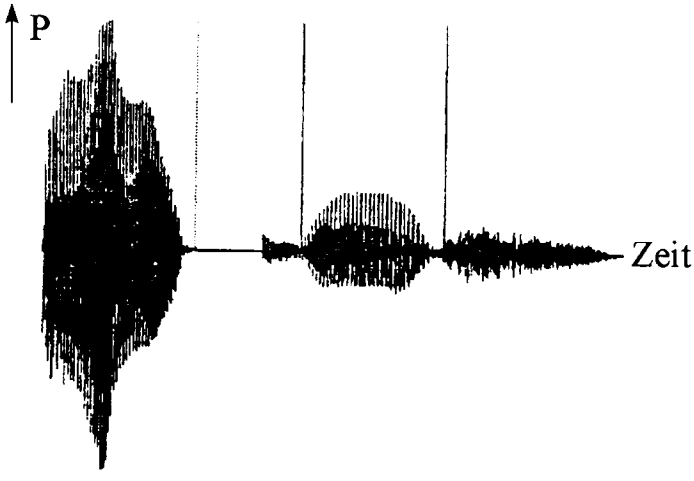

Bild 3: Automatisch ermittelte Lautgrenzen des Wortes 'Autos' mit seinen Lautgrenzen (Schwellwert = 0.98)

Literatur

[1] Föst, S.; Christoph, M.: Untersuchung von Verfahren zur automatischen Segmentierung von gesprochenen Lauten. TU Berlin: Studienarbeit 1989.

[2] Sehhati, Sh.; An interactive tool for segmentation and accoustic-phonetic analysis of speech "SPANEX". Invited paper, IEE 6th International Conference Digital Processing of Signals in Communication, Loughborough, U.K., 2.-6. September 1991.

[3] Sehhati, Sh.; Erstellen von Lautelementebibliotheken unter Verwendung von Phonemclustern auf der Grundlage des LPC-Sprachsyntheseverfahrens. Dissertation, TU Berlin Juli 1995.

[4] Siebs, Th.; Deutsche Aussprache, Reine und gemäßigte Hochlautung mit Aussprachewörterbuch. Berlin : Walter de Gruyter & Co., 1969.

[5] van Hemert, J. P.: Automatic Diphone Preparation. IPO, Annual Progress Report 20, S. 23-32, Eindhoven, 1985.

[6] van Hemert, J. P.: Automatic Segmentation of Speech into diphones, Philips Technical Review, vol. 43, No. 9, Sept. 1987.