Eine Segmentierung von Sprachsignalen mit Hidden-Markov-Modellen
T. KUHN, V. FISCHER, S. KUNZMANN, E. NÖTH
(Lehrstuhl für Informatik 5 (Mustererkennung), Friedrich-Alexander-Universität Erlangen-Nürnberg)

1. Einführung

2. Anforderungen an ein flexibles Segmentierungsverfahren
Die Ausgangsbasis für die Segmentierung eines Sprachsignals in lautliche Einheiten bildet das in konsekutive Zeitscheiben mit fester zeitlicher Länge (Frames) partitionierte Sprachsignal. Im Akustik-Phonetik-Modul (APM, [3]) wurde die zeitliche Dauer eines Frames deutlich kürzer als ein Laut gewählt, so daß von der Annahme ausgegangen werden kann, daß sich innerhalb eines Frames die Artikulation des gesprochenen Lautes nicht ändert. Die Zeitfenster mit 12,8 Millisekunden Dauer sind allerdings genügend groß um lauttypische Information zu enthalten. Mit einem Normalverteilungsklassifikator werden im APM 49 Lautkomponentenklassen unterschieden und jeder Zeitscheibe werden die fünf Klassen mit der besten Bewertung zugeordnet. Die durchschnittliche Lautdauer in der EVAR-Stichprobe von 6,5 Frames und eine sprecherunabhängig auf Lautkomponentenebene erzielte Erkennungsrate von 78 Prozent bei Betrachten der ersten fünf Lautkomponentenalternativen verdeutlichen die Gültigkeit der beiden Annahmen.

Um Untersuchungen zum Einfluß der Segmentierung in lautliche Einheiten auf die Erkennungsraten auf Wortebene zu ermöglichen, wurde ein neuartiges Segmentierungsverfahren nach Wortuntereinheiten (z. B. Laute) entwickelt, welches die Berücksichtigung der folgenden Anforderungen ermöglicht:


1 In der ersten Alternative werden ca. 40 Prozent der Lautkomponenten richtig erkannt, für Einzelheiten siehe [3].

Fortschritte der Akustik DAGA '90
Der Segmentierer soll mit unterschiedlichen Eingabedaten arbeiten können. Neben einer symbolischen Beschreibung des Sprachsignals (klassifizierte Frames) sollte das Signal auch durch Merkmalektoren repräsentiert werden können.

Der Segmentierer soll ohne grundlegende Veränderungen auch bei Umstellung der Systemumgebung eingesetzt werden können (Variation der Wortuntereinträge, Veränderung der Abtastrate, neue Merkmalsätze, unterschiedliche Dauer der Zeitscheiben). Dies erfordert, daß ein leichtes Modellieren der zu unterscheidenden Wortuntereinträge möglich sein muß. Die Anpassung an die neue Erkennungsumgebung sollte automatisch durchführbar sein.


3. Charakterisierung des Segmentierungs-/Klassifikationsverfahrens
Um den beschriebenen Anforderungen gerecht zu werden, wurde das Verfahren in die folgenden vier Phasen zerlegt:

1) Bestimme eine Initialzerlegung des Sprachsignals in homogene Bereiche auf der Basis von äquidistanten Zeitscheiben. Das Signal kann dabei sowohl durch Merkmalektoren als auch durch klassifizierte Zeitscheiben gegeben sein.

2) Leite aus dieser Zerlegung ein Segmentnetz ab, indem für alle gefundenen Grenzen (potentielle Anfangspunkte einer Wortuntereinheit) ein Menge von alternativen Endpunkten betrachtet werden.

3) Klassifiziere das Segmentnetz anhand eines vorgegebenen Lautinventars.

4) Führe eine Pfadsuche auf dem bewerteten Segmentgraph durch, um die optimale lineare Segmentierung des Signals zu finden.


1) Die homogenen Bereiche der Initialzerlegung ergeben sich aufgrund der 'gleichen Klassifikatorentscheidung bei konsekutiven Zeitscheiben'. Die Initialzerlegung ergibt sich damit aus den Bereichen,
in welchen sich die erste Alternative im Strom der Lautkomponentenhypothesen nicht ändert. Um die Zahl der potentiellen Lautgrenzen zu reduzieren werden einfache Glättungsoperationen angewendet, falls der Bereich auf einen Klassifikationsfehler innerhalb eines Lautes deuten. Dies sind z. B. Segmente der Länge 1 mit gleicher linker und rechter Klassenbezeichnung.

2) Da sich vor allem an Lautübergängen die Klassifikatorentscheidung häufig ändert und Plosive sowie Diphthonge aus unterschiedlichen Lautkomponenten bestehen, werden für jeden Startpunkt der Initialzerlegung alternative Endpunkte betrachtet. Die Menge der alternativen Endpunkte ist durch die initiale Segmentierung gegeben. Um die Größe des Segmentnetzes zu beschränken, wird die maximale Länge eines Lautes (d. h. die mögliche Zeithöhe einer Kante des Segmentnetzes) beschränkt. Initialsegmente mit einer größeren Dauer als die maximal zugelassene Lautlänge werden allerdings nicht aufgetrennt.


4) Mit Hilfe eines Graphsuchverfahrens (A*-Algorithmus) wird im bewerteten Segmentnetz derjenige Pfad bestimmt, dessen globale Ähnlichkeit zum Sprachsignal am größten ist. Die Kosten eines Pfades ergeben sich aus dem negativen Logarithmus der Ähnlichkeitsbewertungen der Kanten, wobei die stochastische Unabhängigkeit der Bewertungen von den Vorgangerkanten angenommen wird.


4. **Experimentelle Beurteilung des Segmentierers**

Die Beurteilung des Segmentierungs-/Klassifikationsverfahrens erfolgt nicht lokal gemäß dem Optimierungsmerkmal *Anzahl korrekt segmentierter und klassifizierter Wortuntereinheiten*, sondern anhand der Erkennungsleistung auf Wortebene. Für die Untersuchungen wurde ein 36 Lautklassen umfassendes Modellinventar verwendet, in dem jeder Laut durch drei elementare Markovmodelle

Die vorliegenden Arbeiten wurden im Rahmen des BMFT-Verbundvorhabens 'Sprachverstehende Systeme' durchgeführt.