Robuste Spracherkennung ausgehend von stimmhaften Abschnitten mit hohem SNR

Hans-Günter Hirsch, Frank Kremer

Hochschule Niederrhein, Institut für Mustererkennung, 47805 Krefeld,
E-Mail:{hans-guenter.hirsch, frank.kremer}@hs-niederrhein.de

Einleitung

Moderne Spracherkennungssysteme liefern unter guten akustischen Aufnahmebedingungen zufriedenstellende Erkennungsraten. Sobald jedoch das zu analysierende Sprachsignal akustische Störungen beinhaltet, verschlechtert sich die Erkennungsrate deutlich im Vergleich zur menschlichen Fähigkeit des Sprachverstehens in gestörten Umgebungen. In der Vergangenheit wurde bereits eine Vielzahl von Ansätzen zur Verbesserung der Erkennungsleistung unter Störeinfluss untersucht. Prinzipiell lassen sich diese Ansätze grob in zwei Kategorien einordnen. Ansätze der ersten Kategorie versuchen die Sprachanalyse so zu gestalten, dass die resultierenden akustischen Merkmale robust sind, d.h. sie sind möglichst unabhängig von den auf das Sprachsignal einwirkenden Störgrößen. Das von ETSI standardisierte Verfahren [1] stellt ein solches Verfahren dar. Die alternative Vorgehensweise besteht in der Adaption der Referenzmuster an die vorhandenen akustischen Bedingungen. Bei vorheriger Kenntnis der Störungsart kann das Erkennungssystem direkt mit Aufzeichnungen gestörter Sprachsignale trainiert werden. Ansonsten ist eine Anpassung der Modelle durch Analyse der aktuellen akustischen Bedingungen möglich.

Mit beiden Ansätzen ist eine deutliche Verbesserung der Wortfehlerraten möglich, allerdings sind diesen Verbesserungen auch Grenzen gesetzt. Aus der Beobachtung des menschlichen Sprachverstehens haben wir einen alternativen Verarbeitungsansatz entwickelt. Menschen können sprachlich auch noch in Umgebungen kommunizieren, in denen der Störeinfluss so stark ist, dass nur noch einzelne Abschnitte des Gesprochenen verstanden werden. In der Regel werden die Abschnitte mit einem hohen SNR noch verstanden. Diese Abschnitte, in denen in der Regel stimmhafte Laute auftreten, werden vom Menschen zur Rekonstruktion der Sprachnachricht verwendet. Der hier vorgestellte alternative Spracherkennungsansatz geht ähnlich vor. Zunächst werden die stimmhaften Laute einer Äußerung automatisch detektiert. Diese Detektionen dienen als Anfangspunkte für die Erkennung, die ausgehend von den detektierten Stellen zeitlich sowohl vorwärts- als auch rückwärtsgewandt durchgeführt wird, bis eine SNR-Schwelle unterschritten ist. Da die detektierten Stellen im Normalfall innerhalb eines Wortes liegen, werden Teilwortmodelle für die „linke“ und die „rechte“ Worthälfte erzeugt. Die Einführung einer SNR-Schwelle bedingt, dass die unsicheren Bereiche mit niedrigem SNR bei der Wahrscheinlichkeitsberechnung nicht berücksichtigt werden.

Der alternative Ansatz wird in den folgenden Kapiteln beschrieben. Es werden auch erste Ergebnisse für die Erkennung gestörter Varianten der einzelnen Ziffern des TIDigit-Datensatzes präsentiert.

Merkmalsextraktion

Als Grundlage des Verfahrens dient eine robuste Merkmalsextraktion [2], mit der die mit 8 kHz abgetasteten Signale verarbeitet werden. Alle 10 ms werden 25 ms lange Segmente des Sprachsignals analysiert. Es wird eine MFCC-Analyse durchgeführt, die um eine adaptive Filterung im Spektralbereich zur Reduktion von stationären Störgeräuschen und eine „blinde“ Kompensation zur Reduzierung des Einflusses unbekannter Frequenzgänge erweitert wurde. Berechnet werden die Cepstralkoeffizienten C_1-C_{12} und die logarithmische Energie logE sowie die zugehörigen Delta und Delta-Delta Koeffizienten. Daraus resultiert ein Merkmalsvektor mit 39 Komponenten.

Detektion stimmhafter Abschnitte

Für das neue Verfahren wird eine automatische Detektion stimmhafter Sprachabschnitte benötigt. Wir verwenden drei Parameter, die zu einem Detektionsergebnis kombiniert werden.

Der erste Parameter basiert auf einer Analyse des zeitlichen Verlaufs der logarithmischen Energie logE, die bereits bei der Merkmalsextraktion berechnet wird. Nach einer Glättung werden lokale Maxima des Verlaufs berechnet. Falls diese Maxima einen Schwellwert übersteigen, sind sie Kandidaten für stimmhafte Laute. Sofern zwei Maxima innerhalb kurzer Zeit auftreten, werden diese zu einem zusammengefasst.

Der zweite Parameter ist ein Maß für die Stimmhaftigkeit. Der Signalverlauf stimmhafter Laute ist durch das periodische Öffnen und Schließen der Stimmbänder näherungsweise periodisch, wo hingegen der Verlauf stimmloser Lauter rauschähnlich ist. Um diese Eigenschaft zur Detektion auszunutzen, wird zunächst aus dem DFT-Spektrum das Cepstrum bestimmt. Mit Hilfe der Bestimmung eines Maximums im Cepstrum kann auf die Grundfrequenz der Sprache und damit auf die Periodendauer geschlossen werden. Die Korrelation von jeweils fünf aufeinanderfolgenden Perioden ergibt ein Maß für die Periodizität eines Abschnitts. Ein hoher Wert deutet auf einen periodischen Signalverlauf und damit einen stimmhaften Laut hin. Daher wird der Verlauf dieses Maßes ähnlich der logarithmischen Energie auf Maxima untersucht.

Der dritte Parameter ist ein Wahrscheinlichkeitsmaß. Für jeden einzelnen Merkmalsvektor wird die Emissionswahrscheinlichkeit für den mittleren Zustand der HMMs aller Lautmodelle berechnet. Die Lautmodelle werden zu den Klassen stimmhaft und stimmlos gruppiert. Berechnet wird die mittlere Wahrscheinlichkeit der fünf besten Laute beider Klassen. Als Parameter dient der

zeitliche Verlauf der Differenz der mittleren Wahrscheinlichkeiten beider Klassen. Bereiche in denen die Wahrscheinlichkeit für die stimmhafte Klasse für einige Zeit über der für die stimmlose Klasse liegt, werden entsprechend als stimmhaft betrachtet.

Die drei erwähnten Parameter werden in der Art kombiniert, dass ein Abschnitt als stimmhaft zu betrachten ist, sofern mindestens zwei der Parameter dies anzeigen. Diese Vorgehensweise zur Detektion stimmhafter Laute wird in [3] evaluiert und auch genauer vorgestellt. Für die Entwicklung des Erkennungssystems muss beachtet werden, dass die Detektion möglicherweise nicht alle stimmhaften Laute detektiert und Fehldetektionen auftreten können.

Training

Wie bereits erwähnt ist ein Training von Teilwortmodellen für die linke und rechte Worthälfte erforderlich. Der Zeitpunkt für die Teilung wird durch die beschriebenen Detektionsroutinen festgelegt und wird in Abbildung 1 beispielhaft durch den rot markierten Merkmalsvektor veranschaulicht.

Alle Vektoren vor dem roten Vektor werden für das linke Modell verwendet, alle Vektoren danach für das rechte Modell. Für das Training werden die TIDigit-Trainingsdaten (ca. 8600 Äußerungen) und die entsprechenden Labelinformationen genutzt. Dabei wird die Anzahl der Zustände eines Modells abhängig von der Anzahl der im Mittel repräsentierten Merkmalsvektoren gewählt. Um eine gute Vergleichbarkeit mit Referenzexperimenten zu gewährleisten, werden gemittelt über alle Modelle ca. 16 Zustände pro Wort verwendet. Ebenso aus dem Grund der guten Vergleichbarkeit werden zwei Mischverteilungen verwendet. Eine spezielle Betrachtung ist bei den Wörtern mit mehreren stimmhaften Lauten („seven, „zero“) erforderlich. Bei diesen Wörtern (und auch bei „four“ und „five“) werden je zwei Teilmodelle für jeden der beiden detektierten stimmhaften Abschnitte erzeugt. Die Gesamtzahl der HMMs für die elf Ziffern (1-9 und „zero“/„ow“) beträgt entsprechend 60, da bei sieben Wörtern je zwei Teilmodelle und bei vier Wörtern je vier Teilmodelle nach Geschlechtern getrennt erzeugt werden.

Modifizierte Erkennung

Die Erkennung gestaltet sich so, dass zwei unabhängige Erkennungsprozesse für die beiden Worthälften durchgeführt werden. Der Startzeitpunkt für die Erkennung wird durch die Detektion festgelegt. Die eingeführte SNR-Schwelle führt dazu, dass die Merkmalsvektoren des Wortanfangs und Wortendes bei der Erkennung häufig nicht berücksichtigt werden können. Aus diesem Grund kann die Wahrscheinlichkeit für ein Wort nicht wie üblich im letzten HMM-Zustand des Wortmodells bestimmt werden. Die Berechnung muss in einem Zustand, der bei der jeweiligen Anzahl der Merkmalsvektoren sinnvoll ist, durchgeführt werden. Die Gesamtwahrscheinlichkeit und damit das Erkennungsergebnis werden durch Addition der logarithmischen Teilwahrscheinlichkeiten bestimmt.

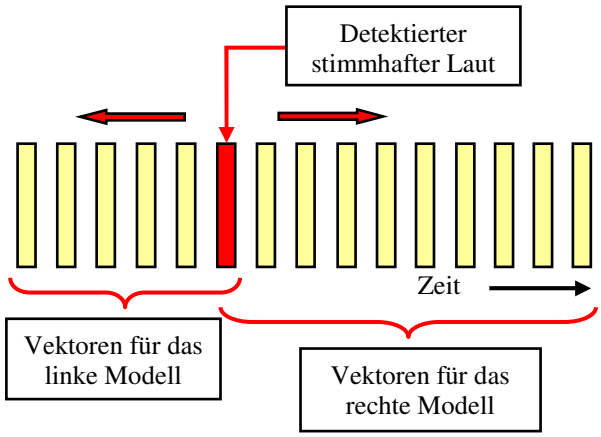

Abbildung 1: Aufteilung der Merkmalsvektoren

Ergebnisse

Betrachtet werden Experimente mit dem Aurora-5 Datensatz [4], der auf der Datenbank TIDigit basiert. Den ungestörten Aufnahmen wurden dabei in fahrenden Autos (car noise) bzw. in unterschiedlichen räumlichen Bedingungen (interior noise) aufgezeichnete Störgeräusche überlagert. Für jedes Störungsszenario existieren 2486 gesprochene Einzelziffern als Testdaten. Zum Vergleich der Erkennungsergebnisse wurden die Tests auch mit einem üblichen HMM-Ganzworterkenner, der dieselbe Merkmalsextraktion nutzt, durchgeführt. Tabelle 1 können die erzielten Ergebnisse entnommen werden.

Die Ergebnisse sind bisher in den meisten Experimenten der Vergleichserkennung unterlegen. Zu beachten ist auch, dass das Erkennungsproblem auf eine Einzelziffererkennung beschränkt wurde. Für die Entwicklung eines Verbundworterkenners sind weitere Untersuchungen erforderlich. Dennoch zeigen diese Ergebnisse, dass das Verfahren im Prinzip funktioniert und möglicherweise künftig einen Beitrag zur Verbesserung der robusten Spracherkennung liefern kann.

Tabelle 1: Wortfehlerraten in %

Variante	Szenario			
	car15db	car10db	car5db	car0db
Referenz	0,88	1,89	4,47	13,07
Neuer Ansatz	1,25	2,33	4,10	10,12
	int15db	int10db	int5db	int0db
Referenz	1,61	3,22	8,61	25,78
Neuer Ansatz	1,81	3,82	10,46	28,56

Literatur

[1] ETSI standard document, “Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced Front-end feature extraction algorithm, ETSI document ES 202 050 v1.1.3 (2003-11), Nov. 2003

[2] Hirsch, H.G., Kitzig, A. Robust speech recognition by combining a robust feature extraction with an adaptation of HMMs, ITG Fachtagung Sprachkommunikation, 2010

[3] Hirsch, H.G., Kitzig, A., Kremer, F. „Detektion stimmhafter Abschnitte zur robusten Spracherkennung“, Workshop Audiosignal- und Sprachverarbeitung, Koblenz, 2013

[4] Aurora project. http://aurora.hsnr.de, Daten verfügbar unter http://www.elda.org, 2007