Untersuchungen zur Kombination der Erkennungsergebnisse zweier robuster Spracherkennungssysteme

Andreas Kitzig, Hans-Günter Hirsch

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

Einleitung

Durch die Kombination einzelner Spracherkennungssysteme zu einem Gesamtsystem kann die Leistungsfähigkeit des Gesamtsystems gegenüber den einzelnen Systemen unter bestimmten Voraussetzungen verbessert werden. Die Verknüpfung mehrerer Systeme kann an verschiedenen Stellen in der Verarbeitungskette von einer Zusammenfassung der akustischen Merkmale [1], [2] bis hin zu einer Kombination der Erkennungsergebnisse [3] erfolgen. Im Rahmen dieser Untersuchungen wurden die Erkennungsergebnisse zweier Systeme verglichen, von denen das eine auf einer Extraktion robuster akustischer Merkmale (HGH robust [5]), das andere auf einer Adaption der HMMs (HGH adapt [4]) beruht. Dabei stellt man fest, dass sich die Ergebnisse bei Betrachtung verschiedener Störszenarien zu einem gewissen Anteil „orthogonal" zueinander verhalten. Dies bedeutet, dass nur einer der beiden Erkenner ein richtiges Ergebnis liefert. Daraus ergibt sich ein Potential zur Verbesserung der Erkennungsrate des Gesamtsystems. Es wurden zwei Verfahren zur Kombination der Erkennungsergebnisse der beiden robusten Systeme implementiert, die im folgenden Text vorgestellt werden. Durch die Verwendung der Kombinationsansätze konnte eine Verbesserung der Erkennungsraten im Rahmen von Erkennungsexperimenten mit gestörten und ungestörten Sprachdaten der Aurora-5 Datenbank [6] für beide Ansätze aufgezeigt werden.

Aufbau der Verfahren

Die beiden im Rahmen dieser Untersuchungen betrachteten Verfahren zur Verknüpfung zweier parallel betriebener Spracherkennungssysteme basieren auf einer Auswertung der erzielten Erkennungsergebnisse.

Dabei wird bei dem ersten Kombinationsverfahren nicht nur die erkannte Folge von Wörtern bzw. HMMs jedes Erkennungssystems, sondern auch die aus der Erkennung resultierende Zuordnung der Wortmodelle zu bestimmten zeitlichen Abschnitten des Sprachsignals ausgewertet. Für die zeitlichen Abschnitte, bei denen sich abweichende Erkennungsergebnisse einstellen, wird für die beiden Erkenner ermittelt, wessen Ergebnis eine höhere relative Wahrscheinlichkeit besitzt. Dazu werden die erkannten Wörter bzw. HMMs der beiden Systeme gemäß der zeitlichen Reihenfolge ihres Auftretens auf ihre prozentuale zeitliche Überlappung hin untersucht. Werden parallel die gleichen Wörter mit einer hohen prozentualen zeitlichen Überlappung erkannt, so wird dies als ein zuverlässiges Teilendergebnis verwendet. Ist die Überlappung gering und/oder unterscheiden sich die erkannten Wortmodelle, wird das Teilergebnis als unzuverlässig betrachtet und zwischengespeichert. Die Zwischenspeicherung der unzuverlässigen Teilergebnisse geschieht so lange, bis entweder ein sicheres Teilergebnis auftritt oder alle erkannten HMMs verarbeitet wurden. Die prinzipielle Vorgehensweise dieses Verfahrens wird in Abbildung 1 dargestellt.

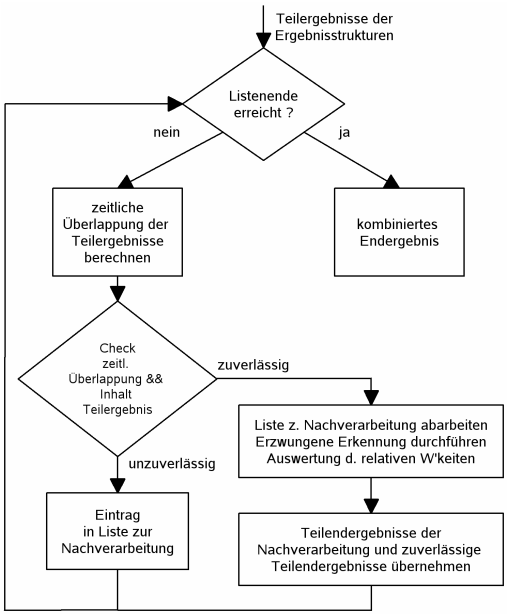

Abbildung 1: Blockschaltbild Kombinationsansatz 1

Für die als unzuverlässig zwischengespeicherten Ergebnisse bzw. für die zugehörigen Signalabschnitte wird bei jedem Erkenner ein weiteres Maß für die Wahrscheinlichkeit des erzielten Ergebnisses bestimmt. Dazu wird eine nochmalige „erzwungene" Erkennung dieser Abschnitte durchgeführt. Hierbei wird dem Erkenner vorgegeben, welche Modelle erkannt werden sollen, eine Erkennung anderer Modelle ist nicht möglich. Mit Erkenner 1 wird das zwischengespeicherte Teilergebnis von Erkenner 1 erneut erkannt. Weiterhin wird mit demselben Erkenner das Teilergebnis von Erkenner 2 verarbeitet. Analog wird mit Erkenner 2 vorgegangen. Es resultieren pro Erkenner zwei erzwungene Erkennungsergebnisse mit entsprechenden Wahrscheinlichkeiten. Für jeden Erkenner wird aus den beiden Wahrscheinlichkeitswerten eine relative Differenz berechnet. Das Erkennungssystem, das die größere relative Differenz liefert, wird als zuverlässiger betrachtet. Anschließend wird das zuverlässigere Teilergebnis als Teilendergebnis verwendet. Anhand dieser Vorgehensweise wird das kombinierte Endergebnis aus allen Teilergebnissen ermittelt.

Das zweite Kombinationsverfahren berücksichtigt im Gegensatz zu dem ersten Verfahren keine Informationen über die zeitliche Zuordnung der erkannten Modelle. Zur

Kombination werden die Folgen erkannter Wortmodelle beider Erkennungssysteme ausgewertet. In Abbildung 2 ist das Blockschaltbild des Verfahrens dargestellt.

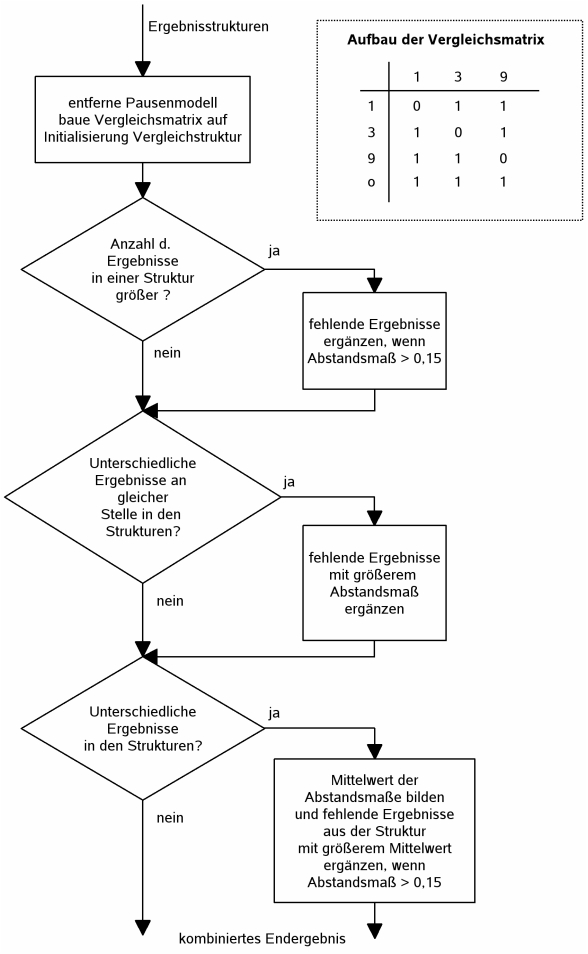

Abbildung 2: Blockschaltbild Kombinationsansatz 2

Zur Auswertung werden die Ergebnisse in eine Vergleichsmatrix eingetragen und verglichen. Gleiche Ergebnisse werden direkt als Teilendergebnis verwendet, für Ergebnisse, die sich inhaltlich und/oder in ihrer Anzahl unterscheiden, wird ein zusätzliches Vertrauensmaß für das jeweilige Ergebnis bestimmt und ausgewertet. Dabei wird überprüft, welches Teilergebnis das höhere Vertrauensmaß aufweist und ob es einen zuvor empirisch ermittelten Schwellwert überschreitet. Wird der Schwellwert bei einem Teilergebnis mit höherem Vertrauensmaß überschritten, ist das Ergebnis zuverlässig und wird als Teilendergebnis verwendet. Zur Bestimmung der Vertrauensmaße wird für jedes erkannte Wort eine Wahrscheinlichkeitsdifferenz berechnet. Die Differenz ergibt sich aus der Betrachtung der Wahrscheinlichkeit des erkannten Wortes und den Wahrscheinlichkeiten, mit denen zu diesem Zeitpunkt das zweitwahrscheinlichste und drittwahrscheinlichste Modell erkannt wurde.

Ergebnisse

Unter Verwendung der beiden Kombinationsverfahren konnten die in Tabelle 1 und Tabelle 2 dargestellten Ergebnisse erzielt werden. Für die Tests wurden aus der Aurora5-Datenbank gestörte Sprachdaten verwendet, die eine Spracheingabe im Freisprechmodus (handsfree bzw. HF) in einer KFZ- und in einer Büroumgebung in Abhängigkeit des Signal-/Rauschleistungsverhältnisses (SNR in dB) simulieren. Jede Testbedingung beinhaltet 8700 englische Sprachäußerungen, die einzelne Ziffern und Ziffernketten beinhalten.

Tabelle 1: Ergebnisse G712CarNoise Handsfree

Testdaten	Systeme			
	HGH adapt	HGH robust	Ansatz 1	Ansatz 2
G712 CarNoise 00dB HF	58,26 %	64,34 %	**65,96 %**	65,53 %
G712 CarNoise 05dB HF	84,74 %	85,43 %	**87,73 %**	87,55 %
G712 CarNoise 10dB HF	95,07 %	94,01 %	**95,93 %**	95,86 %
G712 CarNoise 15dB HF	97,91 %	97,00 %	**98,23 %**	98,09 %

In Tabelle 1 und 2 sind in den Spalten „HGH adapt“ und „HGH robust“ die Worterkennungsraten der robusten Einzelsysteme, in den Spalten Ansatz 1 und Ansatz 2 die kombinierten Ergebnisse des jeweiligen Kombinationsansatzes für die entsprechenden Testumgebungen dargestellt.

Tabelle 2: Ergebnisse Clean / InteriorNoise Handsfree

Testdaten	Systeme			
	HGH adapt	HGH robust	Ansatz 1	Ansatz 2
Clean	99,47 %	99,46 %	99,50 %	**99,51 %**
IntNoise 05dB HF	72,85 %	72,36 %	**76,16 %**	73,46 %
IntNoise 10dB HF	87,54 %	86,10 %	**89,57 %**	87,63 %
IntNoise 15dB HF	93,62 %	92,17 %	**94,76 %**	93,81 %

Die Ergebnisse machen deutlich, dass anhand der vorgestellten Kombinationsansätze eine Steigerung der Worterkennungsrate im Vergleich zu den Einzelsystemen erzielt werden kann. Dabei liefert Ansatz 1 etwas bessere Ergebnisse als Ansatz 2, was auf den komplexeren Aufbau von Ansatz 1 zurückzuführen ist, bei dem neben der erkannten Modellfolge auch die zeitliche Zuordnung ausgewertet wird. Die Steigerung der Erkennungsleistung geht mit einem erheblichen rechnerischen Mehraufwand durch die Verwendung von zwei Erkennungssystemen und deren Kombination einher, was einen Einsatz der Verfahren in Systemen mit eingeschränkten Ressourcen, z.B. in mobilen Endgeräten, erschwert.

Literatur

[1] Lukas Burget. Combination of speech features using smoothed heteroscedastic linear discriminant analysis. ICSLP 004, 2004
[2] Daniel P. W. Ellis. Stream combination before and/or after acoustic model. ICASSP-2000
[3] Jonathan G. Fiscus. A post-processing system to yield reduced word error rates: recognizer output voting error reduction (rover). IEEE Workshop on Automatic Speech Recognition and Understanding, 1997, strony 347-352, 1997
[4] H.G. Hirsch. Automatic speech recognition in adverse acoustic conditions, in Advances in Digital Speech Transmission, John Wiley and sons, 2008
[5] H.G. Hirsch, A. Kitzig: Robust Speech Recognition by Combining a Robust Feature Extraction with an Adaptation of HMMs, 9. ITG Fachtagung Sprachkommunikation, Okt. 2010
[6] Aurora project. http://aurora. hs-niederrhein.de, data available at http://www.elda.org, 2007