Gegenüberstellung von Ansätzen mit robuster Merkmalsextraktion und HMM-Adaptionstechniken

Hans-Günter Hirsch

Institut für Mustererkennung, 47805 Krefeld, E-Mail:hans-guenter.hirsch@hs-niederrhein.de

Einleitung

Zur Erhöhung der Robustheit von Spracherkennungssystemen gegenüber den Veränderungen des Sprachsignals auf Grund der akustischen Umgebung existiert eine Vielzahl von algorithmischen Ansätzen. Dabei stellen die additive Überlagerung von Störgeräuschen und das Auftreten von Nachhall bei einer Spracheingabe im Freisprechmodus zwei akustische Einflüsse dar, die zu einer deutlichen Verschlechterung der Erkennungsraten führen. Die meisten Ansätze lassen sich dabei in zwei Kategorien einordnen.

Zum einen besteht die Möglichkeit die Analyse des Sprachsignals so zu gestalten, dass die extrahierten akustischen Merkmale weitgehend unabhängig von den Veränderungen des Sprachsignals die gleichen oder ähnlichen Werte annehmen. Diese als robuste Merkmalsextraktion bezeichnete Vorgehensweise beinhaltet in der Regel Verarbeitungsschritte zur Reduktion der Störgeräusche und/oder des Nachhalls. Ein Beispiel für ein derartiges Verfahren stellt die von ETSI im Jahr 2003 als Standard veröffentlichte robuste Merkmalsextraktion dar [1]. Dabei findet vor der Cepstralanalyse, auf der die meisten heutigen Spracherkennungssysteme beruhen, eine zweistufige Wiener Filterung zur Reduktion eines stationären Störgeräuschs statt. Des Weiteren wird der Einfluss eines unbekannten Frequenzgangs, z.B. auf Grund des eingesetzten Mikrofons, durch eine „blinde“ Schätzung des Frequenzgangs weitgehend kompensiert. Zur Schätzung werden die aus der Analyse resultierenden Cepstralkoeffizienten mit den Koeffizienten eines mittleren Sprachspektrums verglichen.

Die alternative Möglichkeit zur Erhöhung der Robustheit besteht in einer Adaption der Referenzmuster an die akustischen Bedingungen bei der Spracheingabe. Als Referenzmuster werden in der Regel Hidden Markov Modelle (HMMs) verwendet. Die einfachste Form der Adaption besteht in einem Training oder einer Anpassung der HMM Parameter mit den extrahierten Merkmalen von Sprachsignalen, die ebenfalls in der gestörten Umgebung aufgezeichnet werden. Dies setzt allerdings eine Kenntnis des zu erwartenden Störszenarios voraus. Andernfalls benötigt man eine Schätzung von Parametern, die die Veränderungen auf Grund der aktuellen akustischen Bedingungen beinhalten. Mit den geschätzten Parametern, z.B. dem Spektrum eines stationären Hintergrundgeräuschs oder der Nachhallzeit, werden dann hauptsächlich die Mittelwerte der multivariaten Gaußverteilungen, die das Auftreten der Merkmale in jedem HMM Zustand definieren, adaptiert. Ein Beispiel eines Adaptionsverfahrens ist die unter dem Kürzel MLLR (maximum likelihood linear regression) bekannte Vorgehensweise [2]. Dabei werden die HMM Parameter mit dem Ziel einer Maximierung der Wahrscheinlichkeit, die bei der Generierung der beobachteten Merkmale mit den gegebenen HMMs bestimmt wird, angepasst. MLLR kann sowohl in überwachter Form bei Kenntnis der sprachlichen Inhalte der zur Adaption verwendeten Äußerungen als auch in unüberwachter Form bei Betrieb des Erkennungssystems ohne Kenntnis der jeweiligen Inhalte eingesetzt werden. Der Autor hat die Entwicklung eigener Ansätze zur Erhöhung der Robustheit in beiden Kategorien betrieben, die in den nachfolgenden Abschnitten vorgestellt werden.

Robuste Merkmalsextraktion

Zur Extraktion robuster Merkmale bei Vorhandensein von stationären Störgeräuschen und der spektralen Wichtung auf Grund eines unbekannten Frequenzgangs dient das in Abbildung 1 dargestellte Verarbeitungsschema [3]. Ähnlich wie bei dem von ETSI festgelegten Verfahren erfolgt eine adaptive Filterung der Betragsspektren einer Kurzzeit DFT.

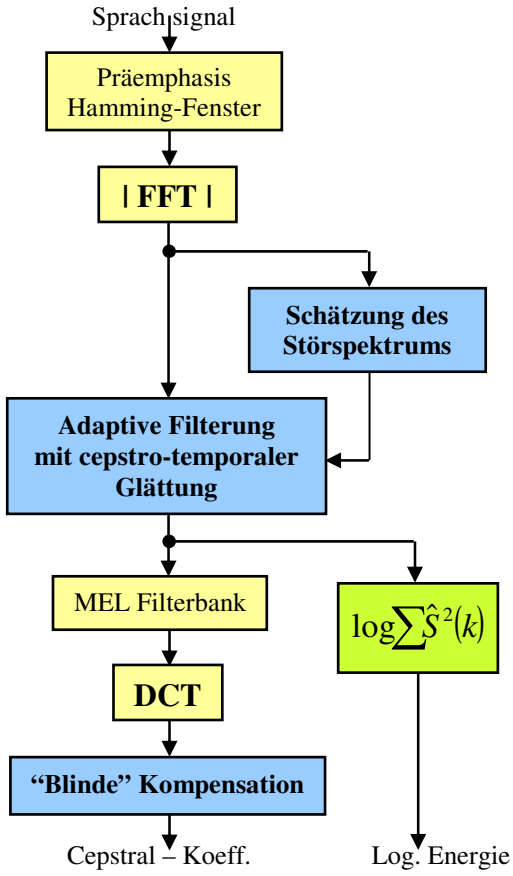

Abbildung 1: Robuste Merkmalsextraktion.

Dabei wird eine cepstrale Glättung der Filtercharakteristik vorgenommen [4], die bei der Zielsetzung einer

Sprachverbesserung zur Reduktion der „musical tones“ eingeführt wurde. Zur Schätzung des Störspektrums wird ein auf der Auswertung der zeitlichen Konturverläufe der DFT Komponenten beruhender Ansatz eingesetzt. Die logarithmierte Kurzzeitenergie wird mit Hilfe des gefilterten DFT Spektrums jedes Segments bestimmt. Von dem ETSI Verfahren wurde der Ansatz der blinden Schätzung und Kompensation eines unbekannten Frequenzgangs übernommen [1].

HMM Adaption

Zur Adaption der HMM Parameter werden verschiedene Parameter geschätzt. Zur Berücksichtigung eines stationären Störgeräuschs wird das MEL Spektrum der Störung aus der Sprachpause zu Beginn einer Äußerung geschätzt. Generell wird die Intention verfolgt, die Adaption so zu gestalten, dass sie in ein in Echtzeit arbeitendes Erkennungssystem integriert werden kann. Die Adaption wird einmalig bei jeder Spracheingabe vorgenommen, wenn der Beginn von Sprache mit Hilfe eines VADs (voice activity detectors) festgestellt wird. Die weiteren Adaptionsparameter sind die Schätzung eines unbekannten Frequenzgangs und der Nachhallzeit für den Fall einer Spracheingabe im Freisprechmodus. Ausgehend von der Annahme, dass sich diese beiden Parameter nicht wesentlich verändern bei Betrachtung aufeinander folgender Spracheingaben, werden diese Parameter nach der vollständigen Verarbeitung und Erkennung einer Eingabe geschätzt und für die nachfolgende sprachliche Äußerung verwendet.

Mit den geschätzten Parametern werden die Mittelwerte der multivariaten Gaußverteilungen aller zur Erkennung verwendeten HMMs adaptiert. Konkret wird der Energieparameter durch die Addition der geschätzten Energie der Hintergrundstörung angepasst. Die Cesptralkoeffizienten werden wie bei der unter dem Begriff PMC (parallel model combination) bekannten Vorgehensweise zurück in den Bereich des linearen MEL Spektrums transformiert. Das lineare MEL Spektrum wird durch Einsatz eines Modells, das die Veränderungen durch den Nachhall berücksichtigt, angepasst. Des Weiteren erfolgen eine Multiplikation des modifizierten Spektrums mit dem geschätzten Frequenzgang sowie eine Addition des geschätzten Störspektrums. Das derart veränderte MEL Spektrum wird wieder in den Cepstralbereich transformiert. Diese Cepstralkoeffizienten werden als modifizierte Mittelwerte der zugehörigen Gaußverteilung benutzt. Zudem wird ein Ansatz zur Adaption der Delta und Delta-Delta Koeffizienten eingebracht. Die Details zur Parameterschätzung und zur Adaption finden sich in [5].

Erkennungsergebnisse

In Abbildung 2 sind beispielhaft die Ergebnisse zur Erkennung gestörter Versionen der englischen Ziffernketten, die als TIDigits Datenbank bekannt ist, dargestellt. Die gestörten Varianten stehen unter der Bezeichnung „Aurora-5“ als eigenständige Datensammlung zur Verfügung [6]. Es werden die Wortfehlerraten angegeben zur Erkennung von Sprachsignalen, bei denen die Aufnahme im Freisprechmodus und die Überlagerung von Störgeräuschen bei einem vorgegebenen SNR simuliert werden. Die Störgeräusche wurden in räumlichen Umgebungen, z.B. in einem Restaurant, einem Einkaufszentrum, etc., aufgenommen. Es werden die Ergebnisse vier verschiedener Erkennungsverfahren verglichen. Die höchsten Fehlerraten treten bei Einsatz einer MEL Cepstralanalyse ohne die in Abbildung 1 eingeführten Verarbeitungsschritte zur Erhöhung der Robustheit und ohne eine Adaption der HMMs auf. Des Weiteren werden die Ergebnisse bei dem alternativen Einsatz der von ETSI standardisierten Merkmalsextraktion oder des zuvor beschriebenen Verfahrens zur Extraktion robuster Merkmale dargestellt. In beiden Fällen stellt sich eine deutliche Reduktion der Fehlerraten ein, wobei die Ergebnisse bei Einsatz des ETSI Verfahrens geringfügig schlechter sind. Das ETSI Verfahren besitzt eine etwas geringere Leistungsfähigkeit, wenn neben Störgeräuschen eine Aufnahme im Freisprechmodus betrachtet wird. Die geringsten Fehlerraten stellen sich ein, wenn man das Verfahren einsetzt, bei dem eine Adaption der HMMs auf die Störumgebung und auf den Nachhall vorgenommen wird. Die relative Verbesserung im Vergleich zu den Verfahren mit robusten Merkmalen fällt bei höherem SNR besser aus.

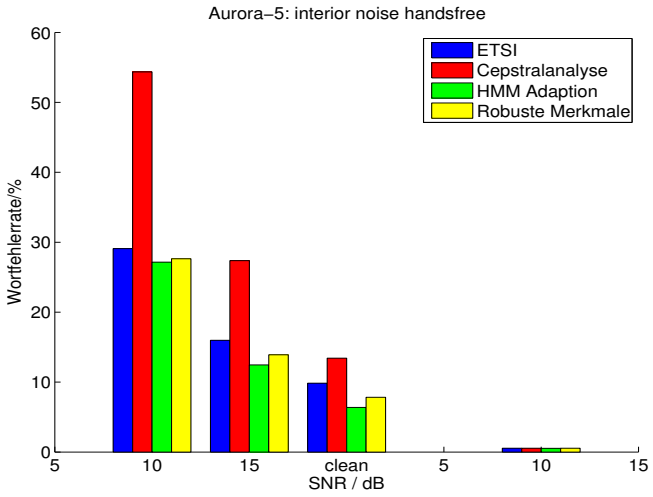

Abbildung 2: Wortfehlerraten bei Überlagerung von Störgeräuschen im Freisprechmodus.

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), 2003

[2] Leggeter, C.J., Woodland, P.C. Maximum Likelihood Linear Regression for speaker adaptation of continuous density HMMs. Computer Speech and Language, Vol.9, 1995

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

[4] Breithaupt, C., Gerkmann, T., Martin, R. Cepstral smoothing of spectral filter gains for speech enhancement without musical noise, IEEE Signal Processing Letters, 2007

[5] Hirsch, H.G., Finster, H. A new approach for the adaptation of HMMs to reverberation and background noise, Speech Communication, Vol.50, 2008

[6] Aurora Projekt, URL: http://aurora.hsnr.de, Daten erhältlich bei http://www.elda.org, 2007