Filterung der Kurzzeit-Energieverläufe in Teilbändern zur Verbesserung der Spracherkennung bei Freisprechen

Andreas Kitzig, Hans-Günter Hirsch

Fachbereich Elektrotechnik und Informatik, Hochschule Niederrhein, 47805 Krefeld
E-Mail: andreas.kitzig@hs-niederrhein.de

http://dnt.kr.hs-niederrhein.de

Zusammenfassung

Bei einem Spracherkennungssystem sinkt die Erkennungsrate erheblich, wenn es in einer räumlichen Umgebung im Freisprechmodus eingesetzt wird. Der Nachhall des Raumes verändert das Sprachsignal, so dass bei einem Training des Systems mit ungestörten Sprachdaten die dabei erzeugten Referenzmuster nicht die akustischen Eigenschaften einer Eingabe im Freisprechmodus beinhalten. Zur Verbesserung der Erkennung wird im Rahmen dieser Untersuchungen ein Modell zur Simulation des Nachhalls im Frequenzbereich betrachtet, da zur Spracherkennung eine Kurzzeit-Spektralanalyse eingesetzt wird. Es werden zwei Einsatzmöglichkeiten dieses Modells aufgezeigt. Zum einen können damit die Referenzmuster an die Hallsituation adaptiert werden. Zum anderen lässt sich aus dem Modell eine inverse Filterung der Energieverläufe in Teilbändern ableiten, um den in einem Sprachsignal enthaltenen Hall zu reduzieren. Diese Filterung kann in die Sprachanalyse integriert werden, die zur Extraktion der relevanten akustischen Merkmale eingesetzt wird. Bei der Durchführung von Erkennungsexperimenten stellte sich bei einer Adaption der Modelle eine deutliche Verbesserung der Erkennungsraten ein.

1 Modell zur Simulation des Nachhalls

Der Einfluss einer Eingabe im Freisprechmodus in einer räumlichen Umgebung kann durch die Faltung eines ungestörten Signals x(t) mit einer Raumimpulsantwort $h_{RIR}(t)$ dargestellt werden. Bei einer Analyse des aus der Faltung resultierenden Signals y(t) im Frequenzbereich stellt man eine Veränderung der Einhüllendenverläufe in den einzelnen Teilbändern fest, die sich anschaulich als künstliche, exponentiell verlaufende Verlängerung der Energieverläufe beschreiben lässt. Houtgast und Steeneken zeigten im Rahmen ihrer Untersuchungen zur Modulationsübertragungsfunktion [1], dass diese Veränderungen als eine Tiefpass-Filterung der Einhüllendenverläufe modelliert werden kann. Daraus lässt sich das in Abb. 1 dargestellte Verfahren zur künstlichen Verhallung von Sprachsignalen ableiten. Die Frequenzanalyse wird dabei mit einer im Bereich der Spracherkennung typischerweise verwendeten Mel-Filterbank vorgenommen. Dabei werden Signalabschnitte mit einer Länge von 25ms Länge mit einer DFT analysiert. Dazu verwendet man bei einer Abtastfrequenz von 8 kHz eine FFT der Länge 256. Aus den resultierenden Betrags-Spektralwerten wird durch eine gewichtete Summation das Mel-Spektrum in 24 Frequenzbändern berechnet. Die Spektralanalyse wird alle 10 ms durchgeführt, so dass am Ausgang 100 Mel-Spektren in einer Sekunde auftreten.

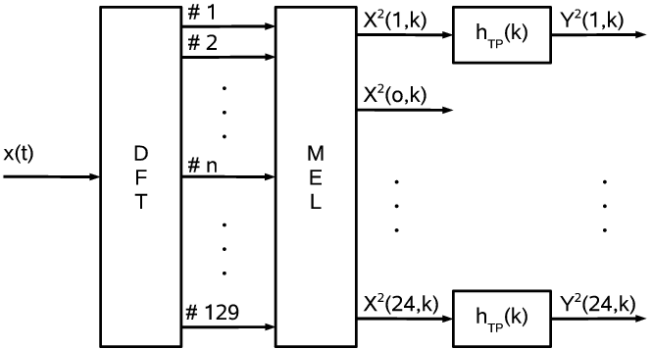

Abbildung 1: Blockschaltbild der Verhallung

Die Faltung der Kurzzeit-Energiespektralwerte $X^2(o,k)$ mit der Impulsantwort $h_{TP}(k)$ eines Tiefpasses lässt sich als FIR Filterung darstellen (mit *o* = Mel-Band-Index und *k* = Frame-Index).

$$Y^2(o,k) = \sum_{i=0}^{N_{Filter}} h_{TP}(i) \cdot X^2(o,k-i) \qquad (1)$$

Anschaulich beschreiben die Werte der Filterkoeffizienten h_{TP} den quantitativen Anteil der Energie eines vorherigen Signalabschnitts, der auf Grund des Nachhalls in einem betrachteten Signalabschnitt auftritt.

Bei Kenntnis der Impulsantwort eines Raumes lassen sich die Filterkoeffizienten durch eine Integration der quadrierten Werte der Impulsantwort über die entsprechend zeitlich angeordneten Signalabschnitte bestimmen. Da man in der Regel ein Erkennungssystem in einem unbekannten Raum einsetzt, dessen Impulsantwort man nicht kennt, kann man einen idealerweise exponentiell verlaufenden Abfall der Impulsantwort annehmen. Die Werte der Impulsantwort lassen sich durch exponentiell gewichtete Werte eines Rauschsignals simulieren. Als einzigen Parameter benötigt man die den exponentiellen Abfall festlegende Nachhallzeit T_{60}, die als frequenzunabhängig angenommen wird.

2 Adaption der Referenzmuster

Zur Spracherkennung werden Hidden Markov Modelle (HMMs) als Referenzmuster verwendet. Das HMM zur Modellierung eines Worts besteht aus 16 Zuständen, wobei die Wahrscheinlichkeit des Verweilens in einem Zustand der

mittleren zeitlichen Länge des durch den Zustand modellierten Sprachabschnitts entspricht. Jeder Zustand beinhaltet die Verteilungsdichtefunktionen eines Satzes akustischer Parameter. Als Parameter werden Cepstralkoeffizienten verwendet, die das Kurzzeitspektrum des Sprachabschnitts beschreiben und die sich durch eine Diskrete Cosinus Transformation (DCT) des logarithmierten Mel Spektrums bestimmen lassen. Zur Anwendung des zuvor beschriebenen Hallmodells werden die Mittelwerte der Cepstralkoeffizienten wieder zurück in den linearen Mel Spektralbereich transformiert, wobei dazu konkret die Mittelwerte der Verteilungsdichtefunktionen herangezogen werden. Die Filterkoeffizienten zur Modifikation der Mel Spektren werden individuell für jedes Modell gemäß der mittleren zeitlichen Längen der Zustände des Modells und mit Hilfe einer geschätzten Nachhallzeit bestimmt. Damit können die adaptierten Mel Spektren bestimmt werden. Diese werden wieder in den Cepstralbereich transformiert, so dass die Mittelwerte der Verteilungsdichtefunktionen adaptiert werden. Die Adaption der Referenzmodelle wird bei jeder Spracheingabe durchgeführt. Die zur Adaption benötigte Schätzung der Nachhallzeit T_{60} wird nach der vorherigen Spracheingabe mit Hilfe einer „maximum likelihood“ Bestimmung vorgenommen. Dazu werden für eine geringfügige Variation der zuvor geschätzten Nachhallzeit jeweils die adaptierten HMMs bestimmt. Es wird die Wahrscheinlichkeit für eine nochmalige Erkennung mit den leicht unterschiedlich adaptierten HMMs berechnet. Die Schätzung der Nachhallzeit wird aus dem Satz adaptierter HMMs abgeleitet, für den die größte Wahrscheinlichkeit berechnet wird.

Nachfolgend werden in Tabelle 1 einige Erkennungsergebnisse für isoliert gesprochene Äußerungen in Form von Wortfehlerraten dargestellt. Als Basis für das Experiment dienten die Sprachdaten der TIDigits Datenbank [3], die mittels zweier Impulsantworten verhallt wurden. Die aus der Faltung mit den Impulsantworten resultierenden Sprachsignale entsprechen den Aufnahmen im Freisprech-modus in einem Büro (office) bzw. einem Wohnzimmer (living).

Tabelle 1: Wortfehlerraten für die Erkennung von Daten aus einer verhallten Umgebung

Experiment	Erkennungs-Modus	
	ohne Adaption	mit Adaption
handsfree office $T_{60}\approx 0{,}4s$	3,49%	1,57%
handsfree living $T_{60}\approx 0{,}6s$	6,64%	2,05%
clean	0,44%	0,44%

Aus den Ergebnissen in Tab. 1 ist deutlich eine Senkung der Wortfehlerrate bei Verwendung der Adaption zu erkennen.

3 Hallreduktion mittels inverser Filterung

Im Gegensatz zur zuvor beschriebenen Adaption der Referenzmuster auf die akustische Umgebung eines Raumes kann man aus dem beschriebenen Hallmodell auch einen Ansatz zur Reduktion des in einem Sprachsignal enthaltenen Halls ableiten. Diese Hallreduktion kann in die zur Erkennung benötigte Sprachanalyse integriert werden. Rechnerisch ist dies mit einem geringeren Aufwand im Vergleich zur Adaption aller HMMs verbunden.
Zur Reduktion des Halls wird aus dem TP-Filter, dass als FIR Filter dargestellt wurde, das inverse HP-Filter abgeleitet, das die gleichen Filterkoeffizienten besitzt und ein IIR Filter darstellt. Die Werte der Filterkoeffizienten lassen sich wie zuvor mit Hilfe einer geschätzten Nachhallzeit bestimmen. Die Integration der inversen HP-Filterung in die Sprachanalyse des Erkennungssystems ist in Abb. 2 dargestellt.

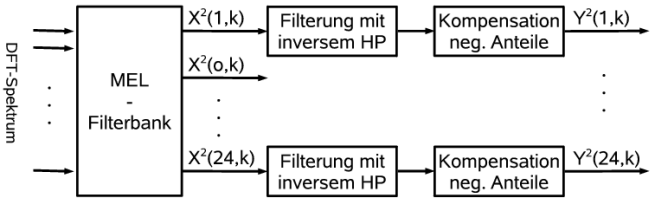

Abbildung 2: Verarbeitungsblock zur Enthallung

Es ergaben sich in allen betrachteten Fällen stabile IIR Filter. Bei einer Filterung der Mel-Spektren von Sprachsignalen, die aus einer Simulation der Eingabe im Freisprechmodus hervorgegangen waren, traten negative Werte bei den Energieverläufen in allen Teilbändern auf. Das macht eine Nachbearbeitung erforderlich. Um die negativen Anteile zu enfernen, wurden zwei verschiedene Ansätze untersucht. Bei dem ersten Ansatz werden alle Energiewerte unterhalb eines kleinen positiven Schwellwerts auf diesen Schwellwert gesetzt. In einem zweiten Ansatz werden negative Werte durch den letzten vorhergehenden positiven Energiewert mit einem zeitlich exponentiell abfallenden Verlaufs dieses positiven Werts ersetzt. Im Rahmen von Erkennungsexperimenten konnten mit dieser Form einer inversen Filterung keine Verbesserungen der Erkennungsrate ermittelt werden, was auf die „unnatürliche“ Modifikation der Mel-Spektren bei Auftreten negativer Energiewerte zurückgeführt werden kann.

4 Fazit

Es wurde ein Verfahren zur Simulation von Nachhall vorgestellt, das auf der Filterung der Einhüllendenverläufe in Teilbändern basiert. Die Erkennungsraten eines Sprach-erkennungssystems konnten deutlich verbessert werden, wenn man diesen Ansatz zur Adaption der Referenzmuster auf die akustische Umgebung eines Raumes einsetzt. Leitet man aus der Hallsimulation eine inverse Filterung zur Reduktion des in einem Sprachsignal enthaltenen Halls ab, so konnte bei einer Integration dieser Filterung in die Sprachanalyse eines Erkennungssystems keine Verbesserung festgestellt werden.

Literatur

[1] H. J. M. Houtgast, T. Steeneken, The modulation transfer function in room acoustics, Acustica, 1973

[2] H.-G. Hirsch, Automatic speech recognition in adverse acoustic conditions, in “Advances in Digital Speech Transmission“, John Wiley & sons, S. 478, 2008.

[3] R.G. Leonard, A Database for speaker-independent digit recognition. Proc. of ICASSP, Vol. 3, 1984