# Verbesserung der Spracherkennung bei Freisprechen durch die Kombination einer robusten Merkmalsextraktion und einer Adaption der Referenzmuster

*Hans-Günter Hirsch, Patrick Pogscheba*

Fachbereich Elektrotechnik und Informatik, Hochschule Niederrhein, 47805 Krefeld
E-Mail: hans-guenter.hirsch@hs-niederrhein.de
Web: http://dnt.kr.hsnr.de/

## Zusammenfassung

Es wird ein Verfahren zur robusten Spracherkennung vorgestellt, das aus der Kombination einer robusten Merkmalsextraktion und einer Adaption der zur Erkennung verwendeten Referenzmuster besteht. Die Extraktion der robusten Merkmale wird mit einem von ETSI standardisierten Verfahren vorgenommen. Damit lassen sich beachtliche Verbesserungen bei der Erkennung von in gestörter Umgebung aufgenommenen Sprachsignalen erzielen. Die Leistungsfähigkeit eines auf diesem Verfahren beruhenden Spracherkennungssystems verschlechtert sich allerdings bei einer Spracheingabe im Freisprechmodus in einer räumlichen Umgebung deutlich. Dabei beeinflusst neben den eventuell vorhandenen Störgeräuschen der Nachhall des Raumes das Sprachsignal. Es wird gezeigt, dass für diese Kombination von Störeinflüssen die Erkennungsraten durch den zusätzlichen Einsatz einer Adaption der Referenzmuster verbessert werden können. Dabei beruht die Adaption auf einer Schätzung der Nachhallzeit des Raumes. Eine derartige Adaption kann bei allen Spracherkennungssystemen eingesetzt werden, die auf einer Extraktion robuster spektraler Merkmale beruhen, um ihre Einsatzfähigkeit auf eine Spracheingabe im Freisprechmodus zu erweitern.

## 1 Einleitung

Einer der wesentlichen Störeinflüsse, der die Erkennungsraten automatischer Spracherkennungssysteme verschlechtert, ist das Vorhandensein von Hintergrundstörungen bei der Spracheingabe. Daher wurden Sprachanalyseverfahren entwickelt, die robuste akustische Merkmale aus einem gestörten Sprachsignal extrahieren. Ein derartiges Verfahren wurde von ETSI im Jahr 2002 als Standard definiert [1].

Die Verwendung eines Freisprechmikrofons in einer räumlichen Umgebung, bei der der Sprecher kein Nahbesprechungsmikrofon zu tragen braucht und zudem die Hände frei zur Bedienung von Geräten zur Verfügung hat, vergrößert die Einsatzmöglichkeiten vieler Spracherkennungssysteme. In diesem Fall wird das Sprachsignal neben der Überlagerung von Störgeräuschen durch die akustischen Eigenschaften des Raumes, die sich als Nachhall bemerkbar machen, beeinflusst. Dabei stellt man fest, dass sich die Erkennungsraten auf Grund des Nachhalls deutlich verschlechtern [2].

Eine Möglichkeit zur Verbesserung der Erkennungsraten besteht in einer Adaption der zur Spracherkennung verwendeten Referenzmuster. Mit Hilfe einer Adaption kann die durch den Nachhall hervorgerufene Veränderung des Sprachsignals in zeitlicher Richtung kompensiert und die Erkennung verbessert werden [3].

Die nachstehend vorgestellten Arbeiten haben zum Ziel, die Kombination einer robusten Merkmalsextraktion und einer Adaption der Referenzmuster zu untersuchen und die Möglichkeiten zur Verbesserung der Erkennungsraten aufzuzeigen.

Im Folgenden wird das standardisierte Verfahren zur Extraktion robuster Merkmale sowie der Ansatz zur Adaption der Referenzmuster auf eine Spracheingabe in verhallter Umgebung vorgestellt. Es werden Ergebnisse zur Erkennung von in verhallter Umgebung aufgenommenen Sprachsignalen präsentiert.

## 2 Robuste Merkmalsextraktion

Das Blockschaltbild des von ETSI standardisierten Verfahrens [1] zur Extraktion robuster akustischer Merkmale ist in Abb. 1 dargestellt. Die Verarbeitung besteht aus einem Verfahren zur Unterdrückung stationärer Störgeräusche, das eine zweistufige Wiener-Filterung mit einer integrierten Schätzung des Störspektrums beinhaltet. Am Ausgang dieses Verarbeitungsblocks zur Störreduktion findet sich ein Signal im Zeitbereich, bei dem die Reduktion stationärer Störgeräusche deutlich hörbar ist.

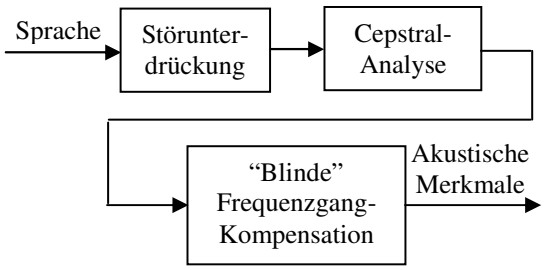


**Abbildung 1:** Blockschaltbild des standardisierten Verfahrens zur Extraktion robuster akustischer Merkmale

Dieses Signal wird im Weiteren einer Kurzzeit-Spektralanalyse unterzogen, wobei die zur Spracherkennung häufig verwendeten Cepstral-Koeffizienten bestimmt werden. Es werden die 13 Cepstral-Koeffizienten niedriger Ordnung einschließlich des Koeffizienten $C_0$ der Ordnung 0 sowie ein Wert für die Kurzzeit-Energie jedes analysierten Sprachsignalabschnitts bestimmt.

Die Cepstral-Koeffizienten werden dann durch Einsatz eines Verfahrens zur blinden Schätzung des Frequenzgangs mit den zugehörigen Cepstral-Koeffizienten

eines „mittleren“ Sprachspektrums verglichen, um den Einfluss unbekannter Frequenzgänge zu kompensieren, die beispielsweise durch das zur Aufnahme eingesetzte Mikrofon oder den Übertragungskanal auftreten können.

Die Effizienz des Verfahrens zur Verbesserung der Erkennungsraten bei Vorhandensein von stationären Hintergrundstörungen und unbekannten Frequenzgängen konnte während des Standardisierungsprozesses als auch in einer Vielzahl späterer Untersuchungen, z.B. [2], aufgezeigt werden.

## 3 Adaption der Referenzmuster

Zur Erkennung wird in vielen Erkennungssystemen eine Modellierung sprachlicher Einheiten, z.B. von Wörtern oder Lauten, mit Hilfe von Hidden-Markov Modellen (HMM) verwendet. Ein HMM besteht aus einer Folge von Zuständen mit bestimmten Übergangswahrscheinlichkeiten zwischen den Zuständen. Der einzelne Zustand eines HMM beinhaltet dabei die akustischen Parameter, die die spektralen und energetischen Eigenschaften eines kurzen Abschnitts des zu modellierenden Sprachsignals beschreiben.

Der Einfluss des Halls eines Raums lässt sich näherungsweise durch eine Impulsantwort h(t) beschreiben, die einen exponentiell abfallenden, zeitlichen Verlauf besitzt, wie es in Abb. 2 dargestellt ist. Dem exponentiellen Abfall von h(t) entspricht ein linearer Abfall der Kurzzeit-Energie, wenn eine stationäre akustsiche Anregung abgeschaltet wird. Die Nachhallzeit definiert dabei die Zeit, bei der die Energie um 60 dB abgefallen ist. Auf Grund dieser exponentiell abfallenden Impulsantwort treten die spektralen und energetischen Merkmale eines Sprachsignalabschnitts in abgeschwächter Form auch zu späteren Zeitpunkten auf und überlagern sich dem Spektrum des späteren Zeitpunkts. Dieser Ansatz der Überlagerung des Spektrums und der Energie eines Signalabschnitts mit den entsprechend abgeschwächten Merkmalen früherer Abschnitte lässt sich auf die Folge von Zuständen eines HMM übertragen und zur Adaption von HMMs auf eine Spracheingabe in verhallter Umgebung einsetzen.

Beispielhaft werden dazu in Abb. 2 die ersten vier Zustände eines HMMs gezeigt. Aus der Übergangswahrscheinlichkeit, die das Verweilen in einem Zustand quantitativ beschreibt, lässt sich die mittlere Dauer des durch den entsprechenden Zustand modellierten Sprachsignalabschnitts bestimmen. Mit Hilfe der Dauer lässt sich der Anteil der Energie, mit dem die charakteristischen Merkmale eines Signalabschnitts zu einem späteren Zeitpunkt auftreten, als Integral über die quadrierte Impulsantwort berechnen, wie es beispielhaft für den dritten Zustand in Abb. 2 dargestellt ist. Zur Adaption wird eine Rücktransformation der Mittelwerte der Cepstralkoeffizienten, die das Kurzzeit-Spektrum eines HMM Zustands definieren, in den Bereich des linearen Mel Spektrums vorgenommen.

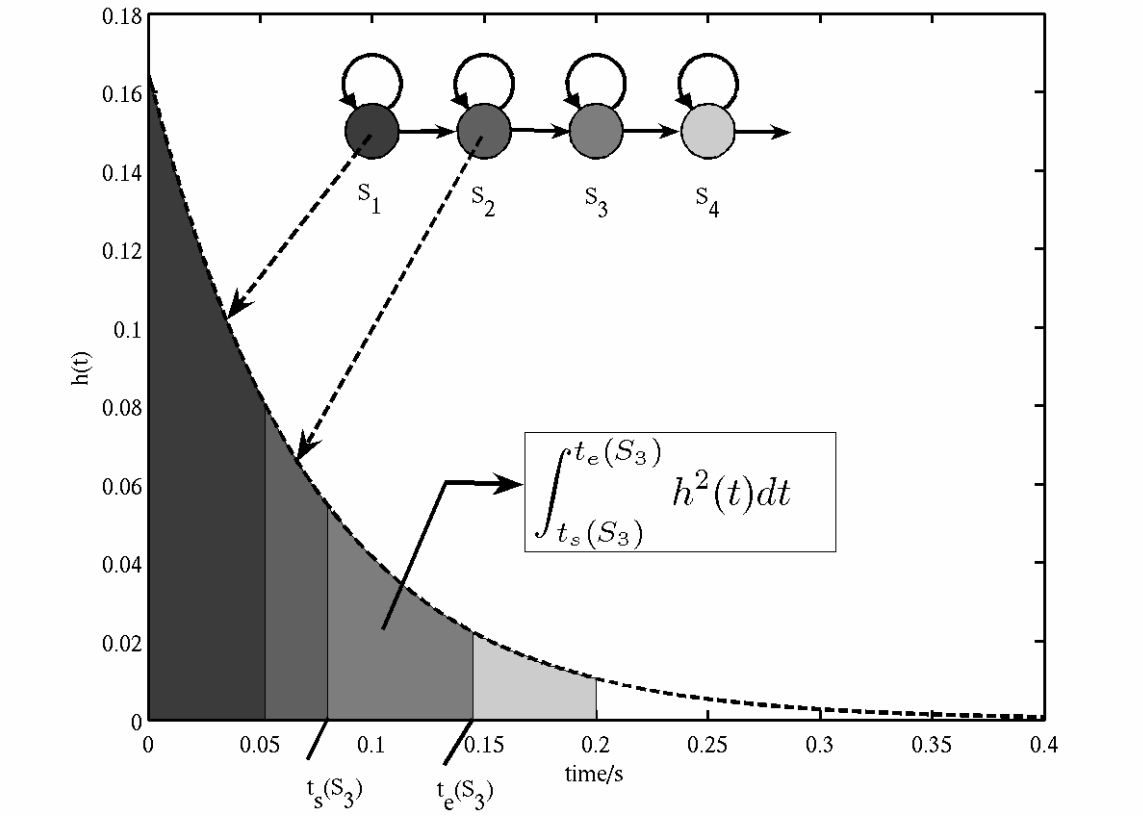


**Abbildung 2:** Exponentieller Nachhallverlauf mit der Energieverteilung auf mehrere HMM Zustände

Das Kurzzeit Leistungsdichtespektrum $|X(S_i)|^2$ eines Zustands $S_i$ wird dabei gemäß der nachstehend beschriebenen, additiven gewichteten Überlagerung von Spektren angepasst:

$$\left|\tilde{X}_k(S_i)\right|^2 = \alpha_{i,i} \cdot \left|X_k(S_i)\right|^2 + \alpha_{i,i-1} \cdot \left|X_k(S_{i-1})\right|^2 + \alpha_{i,i-2} \cdot \left|X_k(S_{i-2})\right|^2 + \cdots = \sum_{j=1}^{i} \alpha_{i,j} \cdot \left|X_k(S_j)\right|^2$$

$$for\ 1 \le k \le NR\_mel$$

Dabei fließen das Spektrum des Zustands i sowie die Spektren aller vorherigen Zustände j=1,2,…,i-1 mit den jeweiligen Wichtungsfaktoren $\alpha_{i,j}$ ein, die den energetischen Anteil der spektralen Merkmale des Zustands j, der zum Zeitpunkt des Zustands i auftritt, festlegen. Der Index k nimmt dabei einen Wert zwischen 1 und der Anzahl NR_mel der Frequenzbänder des Mel Spektrums an.

Die Adaption kann individuell vor jeder neuen Spracheingabe vorgenommen werden. Im einfachsten Fall wird zur Adaption eine Schätzung des Werts der als frequenzunabhängig angenommenen Nachhallzeit benötigt. In diesem Fall besitzen die Wichtungsfaktoren keine Abhängigkeit von der Frequenz. Die Schätzung wird jeweils nach einer Erkennung in Form einer „maximum likelihood“ Bestimmung vorgenommen. Dabei 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.

Die Wirkungsweise der Adaption wird in Abb. 3 veranschaulicht. Darin sind die Spektrogramme von drei HMMs zur Modellierung des englischen Worts „six“ dargestellt. Die Möglichkeiten der Visualisierung der spektralen Merkmale eines HMMs werden in diesem Tagungsband beschrieben [4].
Das HMM in Abb. 3-a wurde durch ein Training mit den ungestörten Aufnahmen der TIDigits Sprachdatenbasis erzeugt. Darin werden die Formantstruktur des Vokalspektrums, die Pause vor dem Plosiv „k“ sowie die höherfrequenten Anteile des Zischlauts am Ende sichtbar. In Abb. 3-b ist das Spektrogramm des HMMs dargestellt, das von den TIDigits Trainingsdaten nach der Simulation einer Aufnahme in einer verhallten Umgebung bestimmt wurde. Darin werden die Nachhallschwänze bei der Betrachtung des Verlaufs der Kurzzeit-Energie in den einzelnen Teilbändern sichtbar.
In Abb. 3-c wird das Spektrogramm der adaptierten Version des HMMs dargestellt, das nach Anwendung des entwickelten Adaptionsverfahrens für einen definierten Wert der Nachhallzeit auf das aus den ungestörten Daten trainierte HMM (Abb. 3-a) erzeugt wurde. Auch in dieser Darstellung werden die Nachhallverläufe in den einzelnen Teilbändern sichtbar. Dies macht deutlich, dass das Adaptionsverfahren zu einem ähnlichen Referenzmuster führt wie ein Training mit Sprachdaten, die im Freisprechmodus aufgenommen wurden.

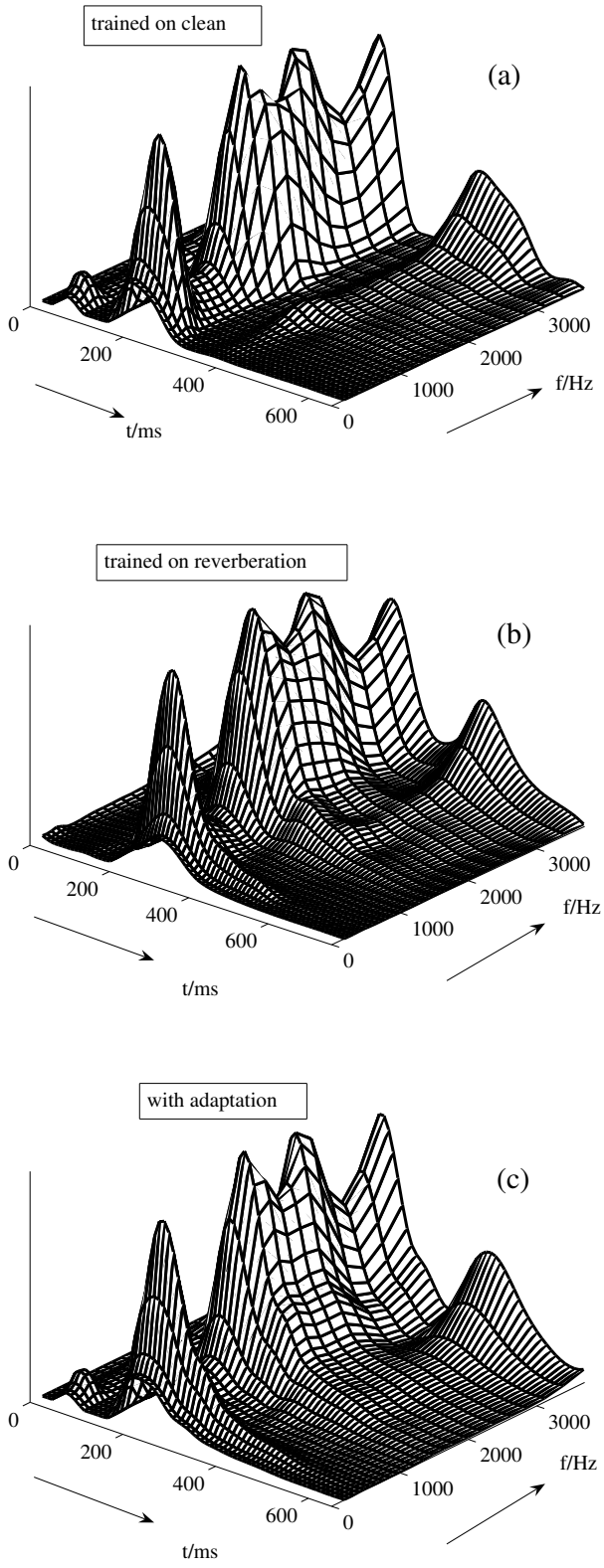


**Abbildung 3:** Spektrogramme drei verschiedener Versionen des HMMs für das Wort „six“

# 4 Erkennungsexperimente

Es wurden einige Erkennungsexperimente mit Hilfe der „Aurora-5“ Datenbasis durchgeführt [5]. Diese Datenbasis wurde unter Verwendung der in der bekannten TIDigits Basis [6] enthaltenen Sprachdaten erzeugt. Es wurde die Aufnahme in einer verhallten räumlichen Umgebung sowie die additive Überlagerung von Störgeräuschen simuliert. Die TIDigits Basis beinhaltet Aufnahmen englischer Ziffernketten. Es stehen etwa 8700 Aufnahmen mit insgesamt etwa 28000 Ziffern zum Training eines Systems sowie die gleiche Anzahl weiterer Aufnahmen für Erkennungstests zur Verfügung. In Abb. 4 sind die Ergebnisse für eine Erkennung der im Freisprechmodus in zwei verschiedenen Räumen aufgenommenen Signale dargestellt.

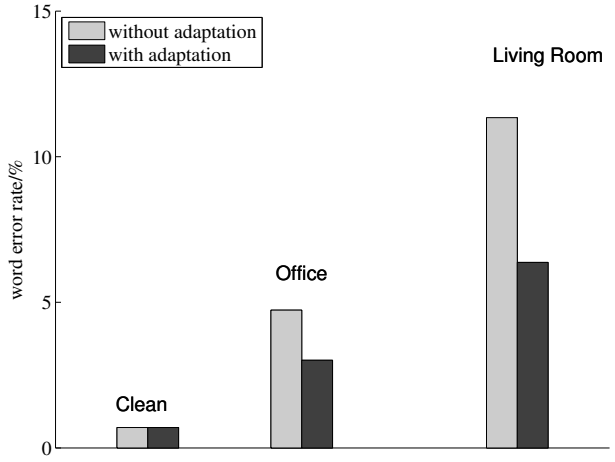


**Abbildung 4:** Wortfehlerraten bei einer Aufnahme in einer verhallten Umgebung

Es werden ein Büroraum mit einer Nachhallzeit von etwa 0.4 s sowie ein Wohnzimmer mit einer Nachhallzeit von etwa 0.6 s betrachtet. Der Abstand des Sprechers zum Mikrofon beträgt in dem Büroraum etwa 1 m und in dem Wohnzimmer etwa 3 bis 4 m. Erwartungsgemäß vergrößert sich die Fehlerrate für die höhere Nachhallzeit. Durch den Einsatz des Adaptionsverfahrens kann die Erkennung deutlich verbessert werden. Die angegebenen prozentualen Werte repräsentieren die erzielten Wortfehlerraten unter Berücksichtigung nicht erkannter als auch zusätzlich erkannter Ziffern bei der Erkennung von Ziffernketten.

In einem weiteren Experiment, in dem die Erkennung der in der Wohnzimmerumgebung aufgenommenen Daten bei dem zusätzlichen Vorhandensein von Störgeräuschen untersucht wird, stellen sich die in Abb. 5 dargestellten Verläufe der Fehlerrate in Abhängigkeit des Signal-/Rauschleistungsverhältnisses (SNR/dB) ein. Bei den Störgeräuschen handelt es sich um die typischen Hintergrundstörungen, die in einer solchen Aufnahmesituation auftreten. Die Störsignale beinhalten vielfach nichtstationäre Abschnitte. Der Einfluss derartiger Störungen kann nur teilweise durch den Einsatz der robusten Merkmals-

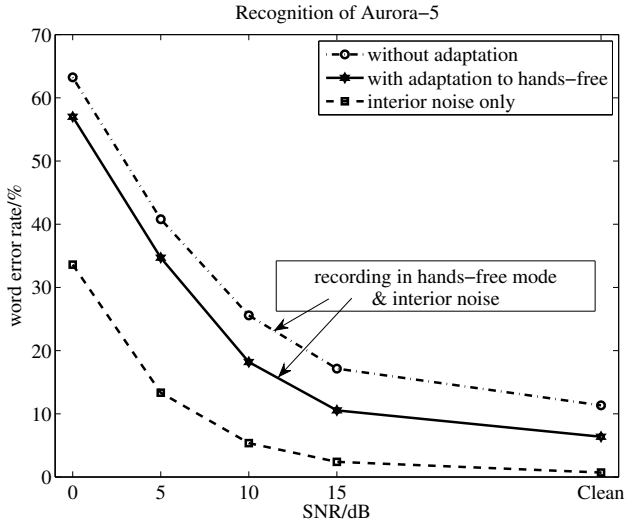


**Abbildung 5:** Wortfehlerraten bei einer Aufnahme in einer gestörten Umgebung

extraktion kompensiert werden. Die Fehlerraten, die ohne eine Aufnahme im Freisprechmodus bei alleinigem Auftreten der Hintergrundstörung erzielt werden, zeigen die Leistungsfähigkeit der robusten Merkmalsextraktion in einer solchen Störsituation. Die Erkennung verschlechtert sich erheblich, wenn neben dem Vorhandensein der Störgeräusche im Hintergrund noch eine Aufnahme im Freisprechmodus betrachtet wird. Die zusätzliche Adaption der Referenzmuster auf den Nachhall des Raumes führt auch in diesem Fall zu einer deutlichen Senkung der Fehlerraten.

Damit wird aufgezeigt, dass die Erkennungsleistung eines Spracherkennungssystems, das auf der Extraktion robuster akustischer Merkmale beruht, durch den zusätzlichen Einsatz einer Adaption der Referenzmuster im Fall einer Aufnahme im Freisprechmodus verbessert werden kann.

# Literatur

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

[2] H.G. Hirsch. Automatic speech recognition in adverse acoustic conditions, in *Advances in Digital Speech Transmission*, Verlag John Wiley and Sons, Jan. 2008

[3] H.G. Hirsch, H. Finster: A new approach for the adaptation of HMMs to reverberation and background noise, *Speech Communication*, Vol.50, S. 244-263, März 2008

[4] H.G. Hirsch, A. Kitzig: Visualisierung der in einem HMM enthaltenen spektralen Merkmale, ITG Fachtagung Sprachkommunikation, Aachen, 2008

[5] Aurora project. http://aurora.hs-niederrhein.de, Daten verfügbar von http://www.elda.org, 2007

[6] R.G. Leonard, A Database for speaker-independent digit recognition. *Proc. of ICASSP*, Vol. 3, p. 42.11., 1984