Visualisierung der in einem HMM enthaltenen spektralen Merkmale

Hans-Günter Hirsch, Andreas Kitzig

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 Visualisierung der in einem Hidden Markov Modell (HMM) enthaltenen spektralen Informationen vorgestellt. Damit besteht die Möglichkeit, neben der mathematischen Darstellung eines HMM als Folge von Zuständen die spektralen Eigenschaften des modellierten sprachlichen Abschnitts zu veranschaulichen und zu analysieren. Mit den aus einer derartigen Analyse gewonnenen Erkenntnissen lassen sich beispielsweise Verfahren zur Verbesserung der Erkennung bei Vorhandensein bestimmter Störeinflüsse ableiten.

1 Einleitung

Zur Modellierung sprachlicher Einheiten wie Wörtern oder Lauten wird bei den meisten heutzutage eingesetzten Systemen zur automatischen Spracherkennung der mathematische Ansatz des Hidden Markov Modells (HMM) verwendet. Dabei wird die jeweilige sprachliche Einheit als eine zeitliche Folge kurzer sprachlicher Abschnitte modelliert. Die Aufeinanderfolge der Abschnitte wird durch den statistisch basierten Ansatz mit Hilfe von Übergangswahrscheinlichkeiten zwischen den Abschnitten beschrieben. Jeder einzelne sprachliche Abschnitt wird durch eine oder mehrere Verteilungsdichtefunktionen der aus der Sprachanalyse resultierenden akustischen Parameter beschrieben.

Insgesamt handelt es sich um einen statistischen Ansatz, der in relativ abstrakter Weise eine sprachliche Einheit mit bestimmten Parametern modelliert. Meist wird ein HMM als die genannte Folge von Abschnitten oder Zuständen dargestellt, ohne die in den Zuständen enthaltenen Merkmale zu visualisieren.

Im folgenden Abschnitt wird eine Vorgehensweise vorgestellt, um die in einem HMM enthaltenen spektralen Merkmale bestimmen und visualisieren zu können. Dies führt zu ähnlichen spektralen Darstellungen, wie sie im Bereich der Analyse von Sprachsignalen an vielen Stellen eingesetzt wird. Solche Visualisierungsmöglichkeiten erleichtern die Analyse des Verhaltens von Erkennungssystemen, um beispielsweise die Problematik in einer bestimmten akustischen Umgebung und das Auftreten von Fehlentscheidungen besser nachvollziehen zu können. Des Weiteren können mit einer solchen graphischen Visualisierung Verfahren zur Verbesserung der Erkennung abgeleitet und beurteilt werden. Im letzten Abschnitt wird gezeigt, wie HMMs mit Hilfe eines Adaptionsverfahrens auf den Einfluss einer Aufnahme im Freisprechmodus in einer gestörten räumlichen Umgebung angepasst werden können.

2 Bestimmung der spektralen Merkmale von HMMs

In den meisten Spracherkennungssystemen werden zur Analyse der spektralen Zusammensetzung eines Signalabschnitts die Cepstral-Koeffizienten niedriger Ordnung extrahiert, mit denen die Einhüllende des Kurzzeit-Spektrums beschrieben wird. Die Cepstral-Koeffizienten werden dabei in der Regel durch eine Diskrete Cosinus Transformation aus dem logarithmierten Mel Spektrum bestimmt. Die spektralen Inhalte eines mit einer Kurzzeit-Cepstralanalyse analysierten Signals lassen sich in Form eines sogenannten Spektrogramms über Frequenz und Zeit visualisieren. Diese Form der spektralen Darstellung wird sehr häufig in der Sprachverarbeitung angewendet, da bestimmte Eigenschaften eines Signals oder eines Verarbeitungsverfahrens besser im Frequenzbereich veranschaulicht werden können.

Zur Modellierung sprachlicher Einheiten wird in der Spracherkennung der Ansatz des HMMs verwendet, wobei man sich dabei auf eine recht abstrakte Darstellung als Folge von Zuständen beschränkt, wie es beispielhaft in Abb.1 für das englische Wort „six“ zu sehen ist.

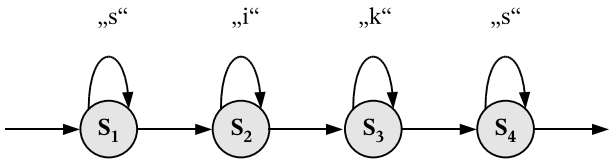

Abbildung 1: Darstellung eines HMMs als Folge von Zuständen

In diesem Beispiel besteht das HMM aus nur 4 Zuständen, die die beteiligte Folge von Lauten repräsentieren. In der praktischen Realisierung wird ein Wort durch eine deutlich größere Anzahl von Zuständen modelliert. Die Übergänge zwischen den Zuständen werden durch Übergangswahrscheinlichkeiten beschrieben.

In den meisten Fällen beschränkt man sich bei den Übergängen zwischen den Zuständen auf die beiden Möglichkeiten des Verweilens in einem Zustand oder des Übergangs in den zeitlich darauf folgenden Zustand. Die beiden Möglichkeiten zum Verlassen eines Zustands werden durch zwei Werte für die zugehörigen Übergangswahrscheinlichkeiten beschrieben, wobei diese Werte bei einem sprecherunabhängigen System in der Trainingsphase aus der statistischen Analyse einer Vielzahl von Äußerungen der zu modellierenden sprachlichen Einheit bestimmt werden. In umgekehrter Weise kann man aus

der bedingten Wahrscheinlichkeit $p(S_t = S_n \mid S_{t-1} = S_n)$ des Verweilens in einem Zustand S_n bei dem zeitlichen Übergang von $t-1$ auf t die mittlere Dauer des Sprachsignalabschnitts abschätzen, der durch diesen HMM Zustand beschrieben wird:

$$dur(S_n) = \frac{1}{1 - p(S_t = S_n \mid S_{t-1} = S_n)} \cdot t_{shift}$$

Dabei entspricht t_{shift} dem Wert für die zeitliche Verschiebung des Analysefensters, der in vielen Systemen einen Wert von etwa 10 ms annimmt.

Der einzelne Zustand eines HMMs versucht den zugehörigen Sprachsignalabschnitt gemäß der in der Sprachanalyse bestimmten Parameter durch eine mehrdimensionale Verteilungsdichtefunktion dieser Parameter zu beschreiben. Dabei wählt man die akustischen Parameter in der Regel so, dass sie als weitgehend statistisch unabhängig voneinander angesehen werden können. Dies ist der wesentliche Grund, weshalb man beispielsweise eine weitere Transformation des Mel-Spektrums in den Cepstralbereich vornimmt. Bei Annahme der statistischen Unabhängigkeit kann man jeden einzelnen akustischen Parameter unabhängig voneinander durch eine oder mehrere eindimensionale Verteilungsdichtefunktionen beschreiben. Dabei verwendet man häufig die Modellierung als Gauß-Verteilung, so dass die Funktion durch zwei Parameter, den Mittelwert und die Varianz, definiert ist.

Zur Visualisierung der spektralen Merkmale eines Zustands wird für jeden Cepstral-Koeffizienten der gewichtete Mittelwert im Fall der Modellierung mit mehreren (NR_mix) Gauß-Verteilungen gebildet, wobei jede Verteilung mit dem Wichtungsfaktor $w(mix_j)$ zum Mittelwert beiträgt.

$$\overline{C}_m = \sum_{j=1}^{NR\_mix} w(mix_j) \cdot C_m(mix_j)$$

Mit Hilfe einer Inversen Diskreten Cosinus Transformation kann man aus den Cepstral-Koeffizienten eines HMM Zustands das zugehörige logarithmische und das lineare Mel Spektrum bestimmen.

$$\{\overline{C}_0, \overline{C}_1, \overline{C}_2, ..., \overline{C}_{NR\_cep}\} \xrightarrow{IDCT}$$
$$\{\log(|\overline{X}_1|), \log(|\overline{X}_2|), ..., \log(|\overline{X}_{NR\_mel}|)\} \xrightarrow{EXP}$$
$$\{|\overline{X}_1|, |\overline{X}_2|, ..., |\overline{X}_{NR\_mel}|\}$$

Durch die zuvor bestimmte mittlere Dauer jedes Sprachsignalabschnitts und die mittleren Mel Spektren besitzt man Kenntnisse über die spektralen Merkmale eines HMMs in Abhängigkeit von Frequenz und Zeit. Der Zeitpunkt $t(S_i)$ des Auftretens eines Zustands S_i wird dabei in der Mitte des zeitlichen Abschnitts, der durch diesen Zustand modelliert wird, angenommen.

$$t(S_i) = \sum_{j=1}^{i-1} dur(S_j) + \frac{dur(S_i)}{2}$$

Aus dem zeitlichen Verlauf der Kurzzeit-Spektralwerte in jedem Band der Mel Filterbank werden mit Hilfe einer Spline Interpolation die Werte des Kurzzeit Spektrums im zeitlichen Raster der Sprachanalyse gewonnen:

$$\{|X_k(t(S_1))|, |X_k(t(S_2))|, |X_k(t(S_3))|, \ldots\}$$
$$\xrightarrow{Spline} \{|X_k(0)|, |X_k(t_{shift})|, |X_k(2 \cdot t_{shift})|, \ldots\}$$

3 Spektrale Visualisierung von HMMs

Die zuvor beschriebene Vorgehensweise zur Visualisierung von HMMs wird dazu verwendet, um mit einer in der Programmierumgebung von Matlab erstellten graphischen Benutzerschnittstelle die spektralen Merkmale zweier HMMs vergleichend betrachten zu können. Ein Beispiel dafür wird in Abb. 2 gegeben, in dem in der linken Hälfte das Spektrogramm des deutschen Worts „zwo“ in zwei- und dreidimensionaler Darstellung zu sehen ist. Das zugehörige HMM wurde aus den Aufnahmen einer selbsterstellten Datenbasis gewonnen. In der Sprachdatenbasis sind unter anderem die deutschen Ziffern enthalten, wobei von jeder Ziffer etwa 3900 Äußerungen von 90 verschiedenen Sprechern existieren. Die Ziffern wurden als Einzelziffern und als Ziffernketten bei Nahbesprechung eines Mikrofons in ungestörter Umgebung gesprochen. Von den Sprachsignalen werden mit einer auf einer Mel Filterbank beruhenden Kurzzeit-Cepstralanalyse die 13 Cepstral-Koeffizienten niedriger Ordnung (C_0 bis C_{12}) sowie ein Kurzzeit Energieparameter bestimmt. Aus den zeitlichen Folgen der akustischen Merkmale werden mit Hilfe des Programmpakets HTK [1] die Parameter eines HMMs ermittelt. Die der Visualisierung in Abb. 2 zu Grunde liegenden HMMs bestehen aus 16 Zuständen. Das Auftreten jedes akustischen Merkmals, das als statistisch unabhängig von dem Auftreten der anderen Merkmale angenommen wird, wird durch eine gewichtete Kombination zweier Gaußverteilungen modelliert. In dem dargestellten Beispiel werden die HMMs aus den Äußerungen aller männlichen Sprecher bestimmt.

In der dreidimensionalen Darstellung ist die Frequenzachse linear skaliert, so dass die höhere Frequenzauflösung bei niedrigeren Frequenzen auf Grund der eingesetzten Mel Filterbank sichtbar ist. Deutlich treten die höherfrequenten Merkmale des Zischlauts am Anfang des Worts hervor. Nachfolgend kann man die niederfrequenten Anteile des Vokals „o“ erkennen. In der rechten Hälfte wird das Spektrogramm des HMMs für das Wort „sechs“ dargestellt. Das HMM wurde ebenfalls aus den Äußerungen der männlichen Sprecher der gleichen Datenbasis gewonnen. Man kann die Formantstruktur des Vokals „e“ sowie die durch den Plosiv „k“ bedingte Pause nach dem Vokal erkennen. Am Ende des Worts wird der höherfrequente Anteil des Zischlauts sichtbar. Mit Hilfe dieser vergleichenden Darstellung werden die spektralen Unterschiede zwischen den beiden Wörtern hervorgehoben. Die visuellen Unterschiede machen die Verwendbarkeit dieser HMMs zur Mustererkennung deutlich.

Die in Abb. 2 dargestellte graphische Benutzer Schnittstelle wird auch im Bereich der Lehre eingesetzt. Damit wird das Verständnis der Struktur von HMMs anschaulich unterstützt.

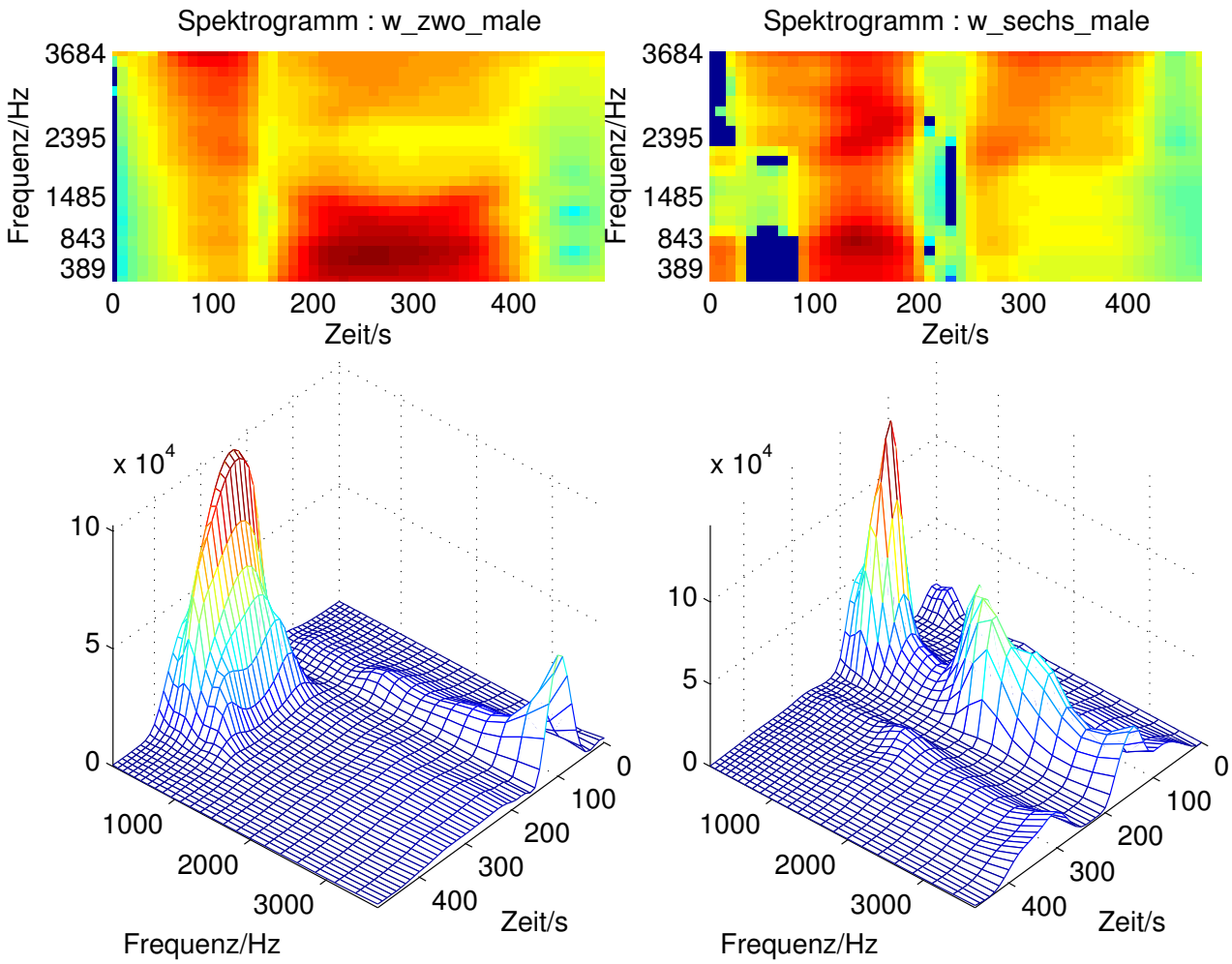

Abbildung 2: HMM-Spektrogramme für die Wörter „zwo“ und „sechs“

4 Analyse mit Hilfe der Visualisierung

Die beschriebenen Visualisierungsmöglichkeiten der spektralen Merkmale in HMMs können zur Analyse des Verhaltens eines Erkennungssystems bei Auftreten bestimmter Störeinflüsse bei der Spracheingabe benutzt werden. Um das Verhalten bei einer Spracheingabe in ungestörter Umgebung mit der Eingabe im Freisprechmodus in einer räumlichen Umgebung zu vergleichen, werden in Abb. 3 zwei Versionen des HMMs für das englische Wort „one“ visualisiert. Das HMM in Abb. 3-a wurde dabei aus den ungestörten Sprachdaten der TIDigits Datenbasis [2] trainiert. In Abb. 3-b ist das entsprechende HMM zu sehen, das aus den Aufnahmen der gleichen TIDigits Daten in einer verhallten Umgebung erzeugt wurde. Dabei wurde die Aufnahme im Freisprechmodus in einer räumlichen Umgebung mit einem vorhandenen Werkzeug [3] zur Simulation verschiedener Bedingungen bei der Spracheingabe simuliert.

Im Vergleich der beiden Spektrogramme wird in Abb. 3-b der Einfluss der Aufnahme in einer räumlichen Umgebung als ein Auftreten der spektralen Merkmale über einen längeren Zeitraum hinweg sichtbar. Dies ist auf den Nachhall des Raumes zurückzuführen, der auf Grund der vielfachen Reflexionen des Schalls in einem Raum zu den exponentiell abfallenden Verläufen der Kurzzeit-Energie in den einzelnen Mel Frequenzbändern führt. Der Vergleich der beiden HMMs macht deutlich, dass bei ausschließlicher Verwendung der aus ungestörten Daten trainierten Referenzmuster die Erkennung von den in einer verhallten Umgebung aufgenommenen Sprachdaten problematisch wird. Tatsächlich führt dies zu einer deutlichen Verschlechterung der Erkennungsraten [3].

Die Autoren benutzen die vorgestellten Visualisierungsmöglichkeiten, um ein Verfahren zur Adaption der Referenzmuster auf eine Aufnahme in einer gestörten, räumlichen Umgebung zu entwickeln und zu optimieren [4]. Das Verhalten des Adaptionsverfahrens wird beispielhaft in Abb. 4 visualisiert. In Abb. 4-a ist das Spektrogramm des HMMs für die deutsche Ziffer „acht“ dargestellt. Dieses HMM wurde aus den Aufnahmen der RVG (Regional Variants of German) Datenbasis [5] gewonnen. Die RVG Daten wurden bei Nahbesprechung eines Mikrofons in einer meist ungestörten Umgebung aufgenommen. Die Formantfrequenzen des Vokals, die höherfrequenten Anteile des Lauts „ch“ sowie die auf die Artikulation des Lauts “t“ zurückzuführende Pause und

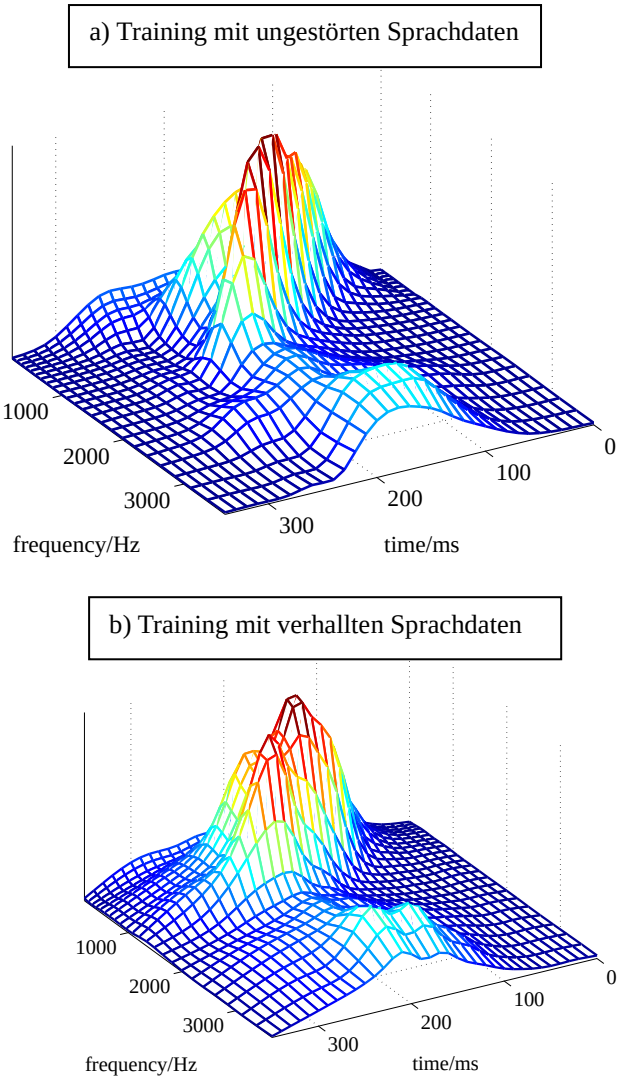

Abbildung 3: HMM-Spektrogramme für das Wort „one“

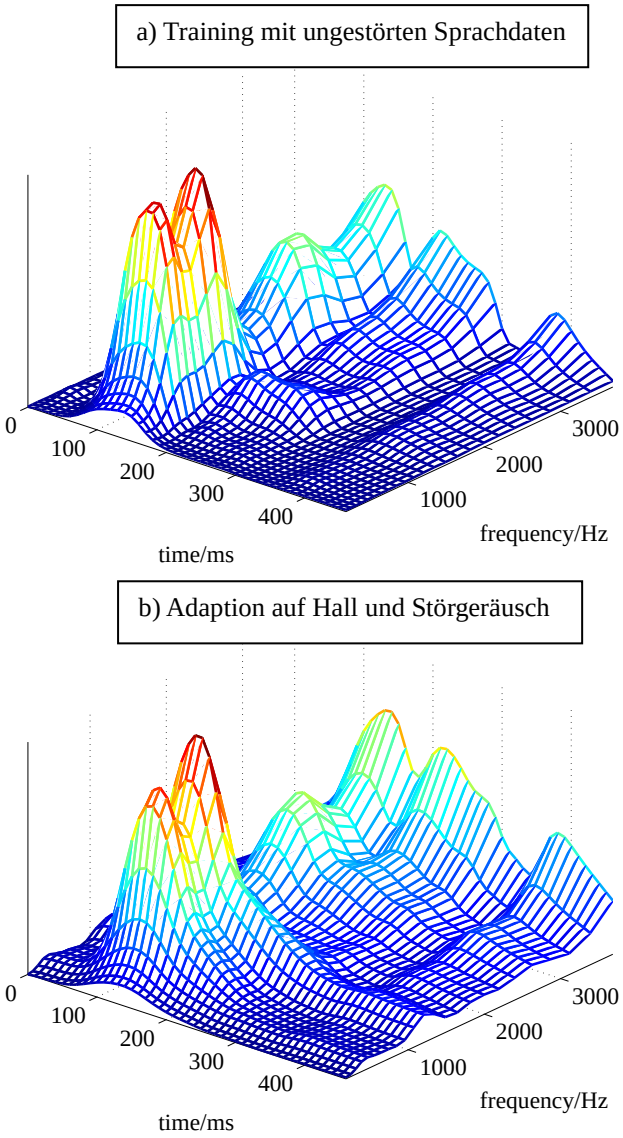

Abbildung 4: HMM-Spektrogramme für das Wort „acht“

das Spektrum des „t“ werden sichtbar.

In Abb. 4-b werden die spektralen Merkmale einer adaptierten Version des HMMs, das Abb. 4-1 zu Grunde liegt, dargestellt. Die Adaption erfolgt dabei individuell bei Detektion des Sprachbeginns einer Äußerung. Das aus der Pause vor der Sprache geschätzte Störspektrum sowie eine aus der Erkennung der vorhergehenden Äußerung gewonnene Schätzung der Nachhallzeit werden zur Adaption des HMMs auf eine Hintergrundstörung und auf den Hall einer räumlichen Umgebung genutzt. In dem gezeigten Beispiel erfolgt die Erkennung einer sprachlichen Äußerung, die im Freisprechmodus in einer räumlichen Umgebung bei Vorhandensein einer typischen Hintergrundstörung aufgenommen wurde. Das geschätzte Störspektrum wird am zeitlichen Ende des HMMs sichtbar. Es führt zu einer Anhebung des gesamten Spektrogramms um das geschätzte Störspektrum. Die Schätzung der Nachhallzeit führt zu den exponentiell abfallenden Nachhallschwänzen bei Betrachtung der Kurzzeit Energieverläufe in den einzelnen Mel Bändern. Der Vergleich mit der Spektraldarstellung eines mit verhallten Sprachdaten trainierten HMMs in Abb. 3-b zeigt, dass das Adaptionsverfahren zu ähnlichen Ergebnissen führt. Dies wird auch durch die deutlich verbesserten Erkennungsraten bei Verwendung dieses Adaptionsverfahrens zur Erkennung gestörter Sprache bestätigt.

Literatur

[1] S. Young et. al.. The HTK book (version 3.3), 2005, http://htk.eng.cam.ac.uk

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

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

[4] 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

[5] Bayerisches Archiv für Sprachsignale, Institut für Phonetik und Sprachverarbeitung, Universität München, http://www.phonetik.uni-muenchen.de/Bas