

REALISIERUNG EINES SPRACHDIALOGSYSTEMS MIT EINER ROBUSTEN SPRACHERKENNUNG

Hans-Günter Hirsch

*Hochschule Niederrhein
hans-guenter.hirsch@hs-niederrhein.de*

Abstract: Es wird ein Sprachdialogsystem vorgestellt, mit dem einfache Dialoge schnell erstellt und modifiziert werden können. Das Dialogsystem basiert auf dem Einsatz eines Verfahrens zur robusten Spracherkennung. Dabei werden die zur Erkennung benötigten Referenzmuster auf die aktuelle Störsituation adaptiert. Die Adaption erfolgt bei jeder Spracheingabe nach der Detektion des Sprachbeginns mit Hilfe eines geschätzten Störspektrums, mit dem ein stationäres Hintergrundgeräusch modelliert wird, sowie mit Hilfe einer geschätzten Frequenzgang-Charakteristik, mit der der zeitinvariante spektrale Einfluss des Mikrofons und des Übertragungswegs beschrieben wird. Neben der Adaption werden weitere akustische Merkmale durch eine robuste Merkmalsextraktion bestimmt. Mit dem Dialogsystem wurden beispielhaft einige Dialoge realisiert, wobei aktuell mit einem Dienst die Verkehrssituation auf den Autobahnen in Nordrhein-Westfalen erfragt werden kann.

1 Einleitung

Einer der Hauptgründe, warum man die automatische Erkennung von Sprache bisher noch relativ selten in praktischen Anwendungen findet, liegt in der Schwierigkeit begründet, eine hinreichend robuste Erkennung zu gewährleisten. Dabei bezieht sich die Robustheit im wesentlichen auf die akustischen Bedingungen bei der Aufnahme und den Einfluss einer eventuell vorhandenen Übertragung der Sprache. Dazu wurden vom Autor schon in der Vergangenheit und werden auch aktuell im Rahmen eines von der DFG geförderten Projekts Untersuchungen angestellt, um den Einfluss aller bei einem praktischen Einsatz der Spracherkennung auftretenden akustischen Randbedingungen festzustellen und Möglichkeiten einer Verbesserung der Spracherkennung in diesen Situationen aufzuzeigen. Konkret werden dabei die folgenden Störeinflüsse betrachtet:

- Freisprechen in einer räumlichen Umgebung
- Stationäre und instationäre Hintergrundstörungen
- Zeitinvariante Frequenzgangveränderungen bei Aufnahme und Übertragung der Sprache
- Sprachcodierung und -decodierung bei einer Übertragung in einem Telefonnetzwerk
- Übertragungsstörungen auf einem Funkkanal

Ein klassisches Beispiel, in der die zuvor genannten Einflüsse auftreten, ist die Freisprechsituation in einem Kraftfahrzeug, in der der Fahrer mit einem telefonbasierten Informationssystem kommuniziert. Bisher konzentrierten sich die meisten Untersuchungen auf das Auftreten stationärer Hintergrundstörungen, ohne die gesamte Vielfalt der Störeinflüsse zu berücksichtigen.

Um die tatsächlich im praktischen Einsatz eines Spracherkennungssystems auftretenden Störeinflüsse feststellen und untersuchen zu können, wurde ein Sprachdialogsystem entwickelt, mit dem beispielsweise einfache telefonbasierte Informationsdienste realisiert werden können. In praktischen Anwendungen treten teilweise Effekte auf, die selbst in großen

Sprachdatenbanken teilweise nicht enthalten sind und damit auch in Simulationsexperimenten keine Berücksichtigung finden. Ein einfaches Beispiel dafür ist das Problem, das Ende einer Spracheingabe zu erkennen, wenn im Hintergrund bei der Spracheingabe noch eine weitere Sprachkommunikation stattfindet. Dieses Problem tritt bei den meisten Sprachdatenbanken gar nicht auf, da bei diesen häufig das Ende einer Aufnahme beschnitten wird. Mit derartigen Erkenntnissen kann man auch noch einen Schritt weitergehen und versuchen, die beobachteten Effekte bei der Erstellung von Datenbanken und Experimenten zur robusten Erkennung einfließen zu lassen.

Ein weiterer Grund für das Entwickeln des Sprachdialogsystems bestand neben der Beobachtung der in praktischen Anwendungen auftretenden Störeinflüsse in der Möglichkeit, damit eigene Ansätze zur robusten Erkennung unmittelbar im praktischen Einsatz testen und ihre Leistungsfähigkeit feststellen zu können.

Im folgenden wird das Dialogsystem zunächst vorgestellt, bevor das darin eingesetzte Verfahren zur robusten Spracherkennung erläutert wird. Im Anschluss werden einige konkret realisierte Informationsdienste vorgestellt, die zur Demonstration des Systems entwickelt wurden.

2 Dialogsystem

Einen Überblick über das gesamte Dialogsystem in seiner Anwendung zur Realisierung telefonbasierter Informationsdienste wird in Abbildung 1 gegeben.

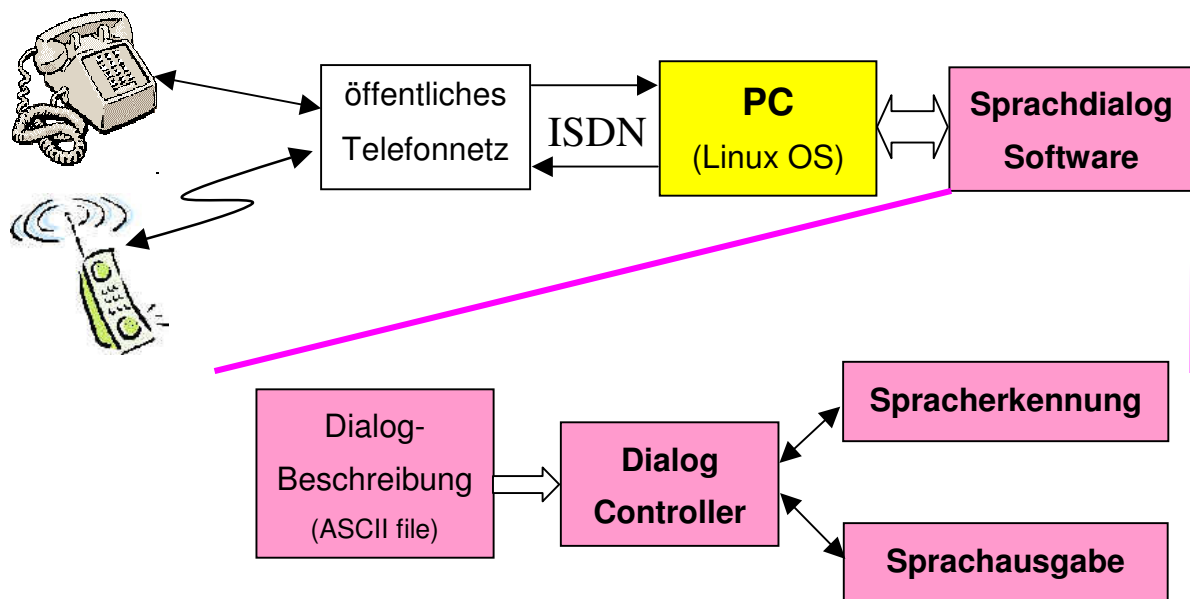


Abbildung 1 – Aufbau des Dialogsystems

Im wesentlichen handelt es sich um eine Software, die auf einem PC unter dem Betriebssystem Linux entwickelt wurde. Die Software kann jedoch relativ leicht auch auf andere Hardwareplattformen und Betriebssystemumgebungen portiert werden, da sie unter C/C++ geschrieben wurde.

In dem PC ist eine herkömmliche ISDN Schnittstellenkarte vorhanden, mit der eine Verbindung zu einer kleinen ISDN Nebenstellenanlage hergestellt wird. Die Nebenstellenanlage ist mit dem öffentliche Telefonnetz verbunden, so dass das Dialogsystem von jedem festen oder mobilen Anschluss aus erreicht werden kann. Außerdem sind an der Nebenstellenanlage direkt ISDN und analoge Telefone angeschlossen, so dass das System auf diese Weise direkt im Labor getestet werden kann.

2.1 Dialog-Controller

Die Software besteht aus einem Dialog-Controller, der wiederum die beiden für den Sprachdialog notwendigen Module zur Sprachausgabe und zur Spracherkennung aufruft. Mit der Entwicklung des Dialog-Controllers wurde das Ziel verfolgt, damit relativ einfache Dialoge entwerfen und schnell ändern zu können. Es war nicht die Intention der Entwicklung eines besonders leistungsfähigen Werkzeugs zur Realisierung komplexer Dialoge vorhanden, um damit sehr leistungsfähige Dienste anbieten zu können oder beispielsweise verschiedene Dialogstrategien vergleichend untersuchen zu können. Vielmehr stand im Vordergrund, mit dem Dialog-Controller die Leistungsfähigkeit der eingesetzten Spracherkennung testen und demonstrieren zu können sowie die im praktischen Einsatz auftretenden Störeinflüsse und Probleme bei der Spracherkennung festzustellen.

Der Dialog kann auf einfache Weise in Textform beschrieben werden. Damit wird beim Start des Controllers der Dialogablauf festgelegt. Den Dialog kann man als ein Netzwerk von Knoten mit einigen Verbindungen zwischen den Knoten beschreiben. In jedem Knoten wird eine bestimmte Funktion, z.B. die Spracherkennung, aufgerufen. In Abhängigkeit des möglichen Rückgabewerts einer Funktion erfolgt dann im Dialogablauf der dem Rückgabewert zugeordnete Übergang in einen anderen Knoten des Dialog-Netzwerks. Beispielhaft wird dies für 2 Knoten in Abbildung 2 dargestellt.

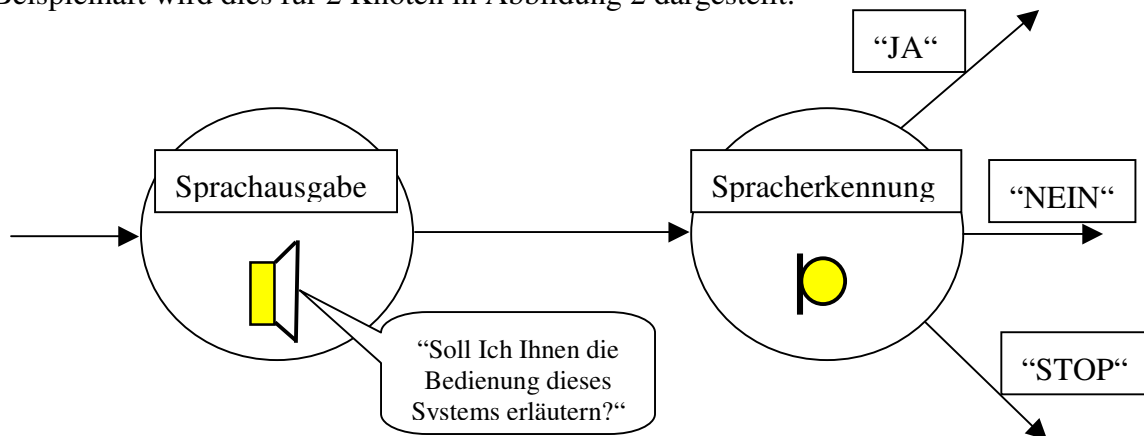


Abbildung 2 – Zwei Knoten eines Dialog-Netzwerks

In diesem Beispiel erfolgt im ersten Knoten eine Sprachausgabe in Form einer Frage. In diesem Fall erfolgt der Übergang in jedem Fall in den darauffolgenden Knoten, in dem die Spracherkennung aktiviert wird, um die Antwort auf die gestellte Frage zu erfassen. In Abhängigkeit des Erkennungsergebnisses erfolgt dann der Übergang in einen weiteren Knoten, um entsprechend in dem Dialog fortzufahren.

Im folgenden werden die Parameter aufgelistet, mit denen ein Knoten vollständig beschrieben wird. Zur Definition eines Knotens wird die Angabe der auszuführenden Aktivität festgelegt, was letztlich zum Aufruf einer bestimmten Funktion führt. Möglicherweise müssen zusätzliche Aktionsparameter angegeben werden, die der Funktion übergeben werden, wie z.B. die Angabe einer auszugebenden sprachlichen Äußerung oder die Angabe, was in einem Erkennungsknoten erkannt werden soll. Des Weiteren muss in Abhängigkeit der möglichen Rückgabewerte der Funktion jedem Rückgabewert der Funktion ein entsprechender Knoten zugeordnet werden, in den dann der Dialog verzweigen soll. Für den Fall, dass kein Rückgabewert angegeben wird oder dass keiner der Rückgabewerte zutrifft, wird noch ein Knoten festgelegt, in dem der Dialog dann fortgeführt wird.

2.2 Sprachein- und ausgabe

Die wichtigsten Module, die zur Realisierung eines Sprachdialogs benötigt werden, sind eine Sprachausgabe und eine Spracherkennung. Die Sprachausgabe kann in der akustischen Wiedergabe einer definierten sprachlichen Äußerung bestehen, was durch die Angabe eines Filenamens festgelegt werden kann. Es kann aber auch die sprachliche Ausgabe eines Erkennungsergebnisses sein, um dieses beispielsweise dem Benutzer als Rückkopplung mitzuteilen. Dazu erfolgt intern eine Abbildung der erkannten Wörter bzw. Modelle auf die entsprechenden Filenamen. Ein "text-to-speech" Modul ist bisher nicht in dem Dialogsystem vorhanden. Die auszugebenden Sprachdaten können als Folge linear quantisierter Abtastwerte oder auch als Folge nichtlinear quantisierter Werte gemäß dem Standard G.711 vorliegen. Erfolgt die Sprachausgabe über ISDN, so werden linear quantisierte Werte auf die zur Übertragung im ISDN notwendige Form der Quantisierung und Codierung gemäß G.711 abgebildet.

Zur Initialisierung der Spracherkennung werden zu Beginn in einem Dialogknoten alle in dem Dialog zur Erkennung benötigten Referenzmuster geladen und intern abgespeichert. In einem Dialogknoten, in dem die Spracherkennung aktiviert wird, wird als Parameter ein Textfile angegeben, in dem die Festlegung der zu erkennenden Modelle bzw. im Fall einer Wortkettenerkennung der Folge zu erkennender Modelle vorgenommen wird. Damit kann dann beispielsweise die Erkennung einzelner Kommandowörter, z.B. nur „ja“ und „nein“, oder aber die Erkennung von Wortketten, z.B. Ziffernketten, festgelegt werden. Zur Erkennung der über ISDN übertragenen Sprache des Anrufers werden die gemäß G.711 quantisierten und codierten Abtastwerte in linear quantisierte Werte umgewandelt, um sie dann mit dem Signalanalyse-Modul der Spracherkennung zu verarbeiten. Zur softwaremäßigen Realisierung der Kommunikation mit der ISDN Karte im PC wird die CAPI [] Schnittstelle verwendet.

Das Spracherkennungs-Modul liefert dem Dialog-Controller die Information über das erkannte Modell bzw. Wort oder die erkannte Folge von Modellen bzw. Wörtern. In Abhängigkeit des erkannten Worts erfolgt dann der in der Dialogbeschreibung festgelegte Übergang in einen weiteren Knoten des Dialogs. Nach einer Erkennung wird die sprachliche Eingabe zusammen mit dem Erkennungsergebnis auf der Festplatte des Rechners gespeichert. Damit können diese Sprachdaten sowohl zum Nachtraining der entsprechenden Referenzmodelle als auch zur Durchführung von Simulationsexperimenten verwendet werden.

Neben den beiden Hauptmodulen zur Sprachaus- und eingabe existiert noch eine Vielzahl weiterer Funktionen. Beispielsweise gibt es einige Funktionen, um in einer tabellenförmigen Zusammenstellung von Daten in Textform nach bestimmten Informationen zu suchen bzw. bestimmte Informationen zu extrahieren. Diese Funktionen werden benötigt, um Informationsdienste zu realisieren, bei denen der Anrufer dem System per Spracherkennung indirekt sozusagen einen Zeilenindex für die Tabelle mitteilt und im Gegenzug weitere, in der Zeile enthaltene Informationen erhält.

3 Spracherkennung

Zur Spracherkennung wird die Weiterentwicklung eines auf der Adaption der Referenzmuster basierenden Ansatzes eingesetzt [1],[2]. Abbildung 3 gibt einen Überblick über das Gesamtsystem.

3.1 Merkmalsextraktion

Aus dem Sprachsignal werden mit einer Kurzzeit-Spektralanalyse 12 Cepstral-Koeffizienten als akustische Merkmale extrahiert. Dabei werden kurze Sprachsegmente mit einer Länge von 25 ms mit Hilfe einer Fourier-Transformation analysiert. Die damit bestimmten Leistungsdichte-Spektralwerte werden gemäß den Charakteristika einer 22 kanaligen MEL-Filterbank zusammengefasst. Die logarithmierten MEL-Spektralwerte werden mit einer Diskreten Cosinus-Transformation (DCT) in den Cepstralbereich transformiert.

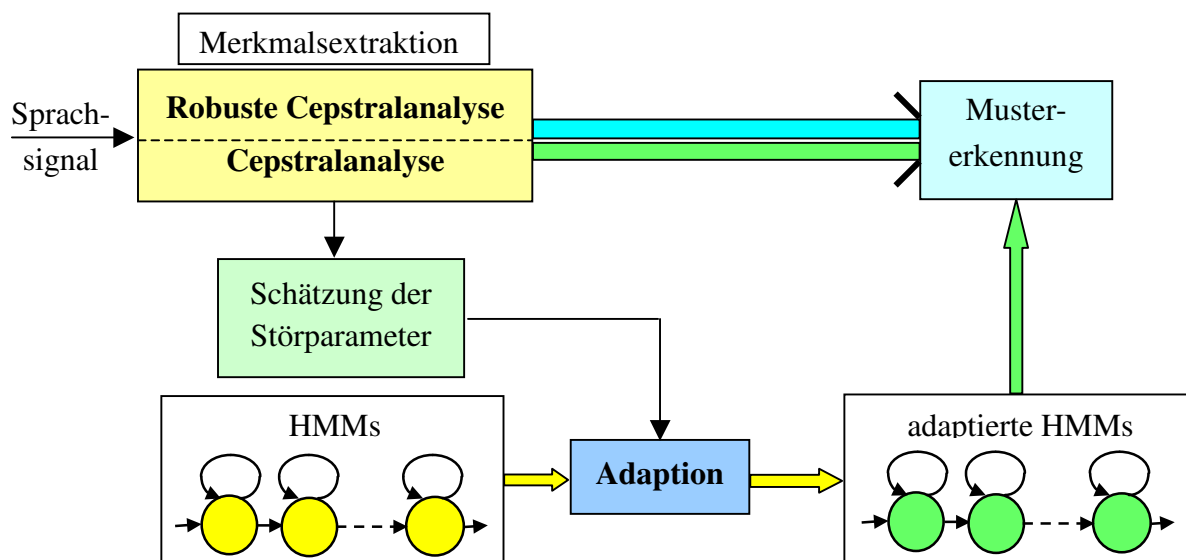


Abbildung 3 – Struktur des Erkennungssystems

Zusätzlich verwendet man meist die sogenannten Delta-Koeffizienten, die sich aus der Ableitung des zeitlichen Verlaufs jedes Cepstral-Koeffizienten bestimmen lassen. Da sich die Adaption der Delta-Koeffizienten auf eine bestimmte Störumgebung als schwierig erweist, wird zur Bestimmung der Delta-Cepstral-Koeffizienten in dem betrachteten Erkennungssystem eine robuste Cepstralanalyse eingesetzt. Diese unterscheidet sich von der zuvor beschriebenen, herkömmlichen Cepstralanalyse durch zwei zusätzliche Verarbeitungsblöcke. Bei den Ausgangswerten der MEL-Filterbank wird zur Reduktion stationärer Hintergrundstörungen eine spektrale Subtraktion des geschätzten Störspektrums vorgenommen. In einem weiteren, unmittelbar darauffolgenden Block wird die Kompensation einer eventuell bei der aktuellen Spracheingabe vorhandenen Frequenzgang-Charakteristik durchgeführt. Dabei werden sowohl das geschätzte Störspektrum als auch die geschätzte Frequenzgang-Charakteristik verwendet, die auch zur Adaption der Referenzmuster benötigt werden. Die Schätzung wird weiter unten erläutert.

Der Wert, der sich aus der Summe der MEL-Spektralwerte ergibt, wird logarithmiert und als weiterer Parameter zur Beschreibung der Kurzzeit-Energie des analysierten Sprachsegments benutzt. Für die zeitlichen Verläufe der ersten 10 Cepstral-Koeffizienten (ohne Berücksichtigung des 0. Koeffizienten) sowie den zeitlichen Verlauf des Energieparameters wird jeweils die Ableitung berechnet, so dass damit 11 Delta-Koeffizienten zur Verfügung stehen. Als weiterer Koeffizient wird von dem zeitlichen Energieverlauf die zweite Ableitung als Delta-Delta-Energiekoeffizient bestimmt. Die 2. Ableitung wird nur für den Energieparameter berechnet, da die Gesamtzahl der Komponenten eines Merkmalsvektors auf 24 beschränkt werden sollte. Ein Merkmalsvektor besteht somit aus

- 12 Cepstral-Koeffizienten ($C_0 - C_{12}$)

- 10 robusten Delta-Cepstral-Koeffizienten ($\Delta C_1 - \Delta C_{10}$)
- 1 robustem Delta-Energiekoeffizient (ΔE)
- 1 robustem Delta-Delta-Energiekoeffizient ($\Delta \Delta E$)

Die Merkmalsvektoren werden im zeitlichen Abstand von 10 ms bestimmt, so dass aus der Signalanalyse 100 Vektoren je Sekunde mit 24 Komponenten je Vektor als Ausgangswerte resultieren.

3.2 Störschätzung

Sowohl zur Realisierung der robusten Cepstralanalyse als auch zur Adaption der Referenzmuster wird eine Schätzung des MEL-Störspektrums benötigt. Dies geschieht bei jeder neuen Spracheingabe durch eine Detektion des Sprachbeginns und eine einfache Mittelung der MEL-Spektren der üblicherweise der Sprache vorausgehenden Sprachpause. Die Detektion des Sprachbeginns wiederum beruht auf einer Untersuchung der Stationarität der MEL Spektralwerte über der Zeit in jedem Teilbank der MEL-Filterbank [2]. Zusätzlich wird auch eine eventuell vorhandene Frequenzgang-Charakteristik geschätzt, die beispielsweise durch das Mikrophon des Anrufers oder den Übertragungsweg verursacht wird. Dies geschieht auf iterative Weise nach jeder erfolgten Spracherkennung durch einen Vergleich der MEL-Spektren des Eingangssignals mit den Spektren der Modellzustände gemäß der wahrscheinlichsten Zuordnung von Merkmalsvektoren zu Modellzuständen [2].

3.3 Adaption der HM-Modelle

Zur Erkennung werden HM-Modelle (Hidden Markov) verwendet, die zuvor mit den Trainingsprogrammen der HTK Software [3] bestimmt wurden. Zum Training der Modelle der deutschen Wörter wurden über Telefon aufgenommene Sprachdaten der SpeechDat Sammlung verwendet. Es werden geschlechtsspezifische Modelle eingesetzt, die aus 18 Zuständen bestehen und bei denen das Auftreten eines akustischen Parameters in einem Zustand durch eine Überlagerung von 2 Gauß-Verteilungen modelliert wird. Das in dem Dialogsystem eingesetzte Erkennungsprogramm kann die im HTK Format abgespeicherten Modelle einlesen und liefert prinzipiell die gleichen Erkennungsergebnisse wie das HTK Erkennungsprogramm.

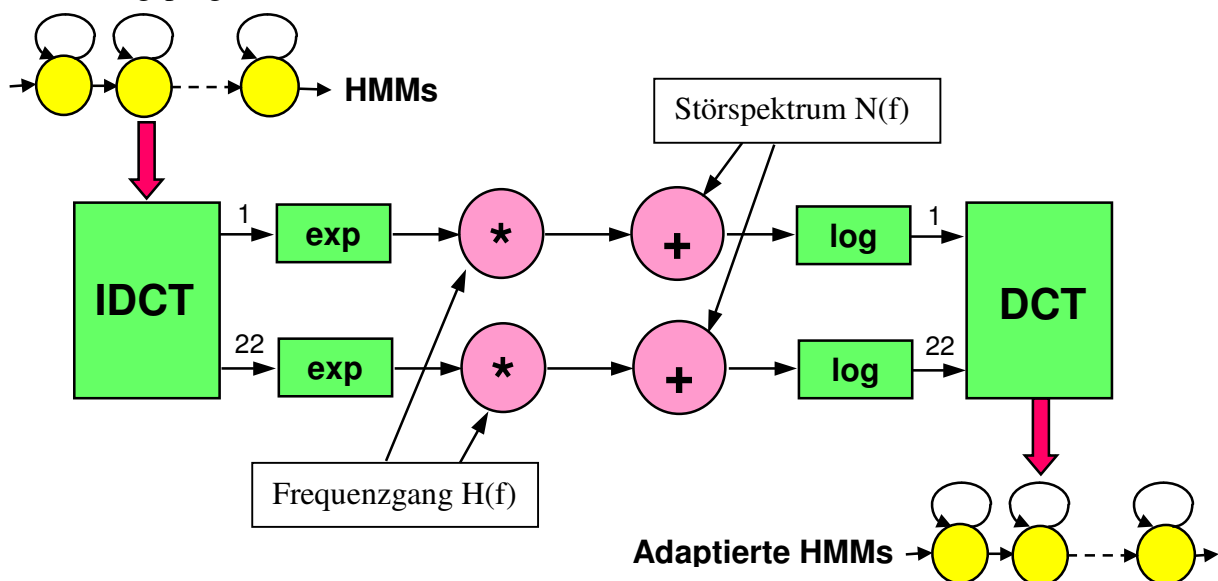


Abbildung 4 – Adaption der Cepstral-Koeffizienten auf die aktuelle Störsituation

Die Adaption der HM-Modelle auf die aktuelle Störumgebung findet bei jeder neuen Spracheingabe nach der Detektion des Sprachbeginns statt. Dazu wird die in Abbildung 4 dargestellte Rücktransformation der Mittelwerte der 12 Cepstral-Koeffizienten jedes Zustands aller HM-Modelle in den linearen MEL Spektralbereich durchgeführt. Im linearen Spektrum kann dann auf einfache Weise die Adaption mit Hilfe der geschätzten Frequenzgang-Charakteristik $H(f)$ und des geschätzten Störspektrums $N(f)$ vorgenommen werden. Abschließend erfolgt wieder eine Transformation in den Cepstralbereich.

Auf diese Weise werden die 12 Cepstral-Koeffizienten aller Zustände aller beteiligten HM-Modelle auf die aktuelle Störumgebung adaptiert. Es erfolgt keine Adaption der zugehörigen Varianzwerte, da dies zum einen rechnerisch sehr aufwendig ist und zum anderen in Erkennungsexperimenten damit nur geringfügige Verbesserungen erzielt wurden [2]. Es existieren auch Ansätze zur näherungsweise Adaption der Delta-Koeffizienten [4], mit denen in Simulationsexperimenten allerdings nur geringfügige Verbesserungen erzielt wurden. Daher werden in dieser Realisierung die durch eine robuste Merkmalsextraktion gewonnenen Delta-Koeffizienten verwendet.

4 Realisierte Dialoge

Mit dem Sprachdialogsystem, das auf der zuvor beschriebenen Spracherkennung beruht, wurden einige beispielhafte Anwendungen realisiert.

Eine erste Demonstration beinhaltete keinen wirklichen Dienst, sondern bestand in einem Frage-Antwort Spiel mit dem Anrufer. Dieser Dialog wurde in englischer Sprache abgewickelt und basierte im wesentlichen auf der Erkennung englischer Ziffernketten. Mit den im Rahmen des praktischen Einsatzes aufgenommenen Sprachdaten wurden im nachhinein einige Simulationsexperimente durchgeführt. Dabei zeigte sich beispielsweise, dass die Wort-Fehlerrate von ca. 8 % ohne Adaption der Referenzen auf ca. 3,5 % durch die Adaption reduziert werden konnte [2].

Aktuell wurden 2 Dialoge in deutscher Sprache erstellt. Bei der ersten Anwendung handelt es sich um eine Notenauskunft für Studenten, in der diese die in einer Klausur erzielte Leistung durch Eingabe ihrer Matrikelnummer erfragen können. Diese Anwendung basiert im wesentlichen auf einer Erkennung deutscher Ziffernketten.

Der zweite Dialog besteht in einer Auskunft über die aktuelle Verkehrssituation auf den Autobahnen in Nordrhein-Westfalen. Dabei wird die aktuelle Verkehrslage durch die regelmäßige Abfrage einer entsprechenden Internet-Seite permanent erfasst und in Form einer Texttabelle dargestellt, die von dem Dialogsystem gelesen und ausgewertet werden kann. Die Spracherkennung ist so gestaltet, dass ein Anrufer beispielsweise die Autobahnnummer 46 als einfache Ziffernkette "vier sechs" oder "a vier sechs" oder auch natürlichsprachlich als "sechs-und-vierzig" oder "a sechs-und-vierzig" sprechen kann. Die Erkennung basiert auch in diesem Fall noch auf einer wortbasierten Erkennung. Die sprachlichen Ausgaben bei diesem Dialog, insbesondere die Namen der Ausfahrten, wurden alle im vorhinein aufgezeichnet. Dieser Dienst ist unter der Telefonnummer 02151-643893 erreichbar, wobei einige Bedienungshinweise im Anhang gegeben werden.

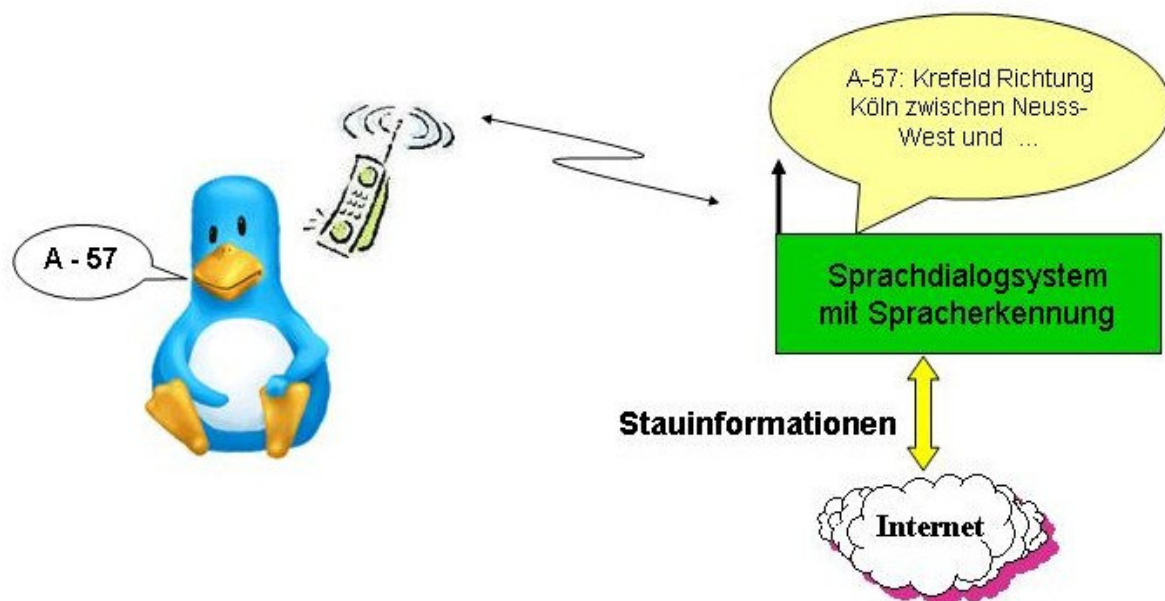
Literatur

- [1] Gales, M.J.F., Young, S.: Robust speech recognition in additive and convolutional noise using parallel model combination. *Computer, Speech and Language* 9, 1995, pp. 289-307
- [2] Hirsch, H.G.: HMM adaptation for applications in telecommunication. *Speech Communication* 34, 2001, pp. 127-139
- [3] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: *The HTK*

book – version 2.2, Entropic, 1999

- [4] Gales, M.J.F.: Nice model-based compensation schemes for robust speech recognition. ESCA workshop on robust speech recognition for unknown communication channels, Pont-a-Mousson, France, pp. 55-64

Bedienungshinweise zum Verkehrsinformationssystem (Tel. 02151 643893)



* Unter der Telefonnummer 02151 643893 (aus dem Ausland +49 2151 643893) ist ein Informationssystem erreichbar, das auf der Basis einer automatischen Spracherkennung Auskunft über die aktuelle Verkehrssituation auf den Autobahnen in Nordrhein-Westfalen erteilt. Zur ständigen Aktualisierung der Verkehrsinformationen wird auf im Internet verfügbare Informationen zurückgegriffen. Bei diesem System handelt es sich um ein Demonstrationsobjekt, mit dem die Ergebnisse der im Labor für Digitale Nachrichtentechnik an der Hochschule Niederrhein stattfindenden wissenschaftlichen Untersuchungen zur robusten Spracherkennung präsentiert werden. Es werden keinerlei kommerzielle Interessen damit verfolgt. Bei der Telefonnummer handelt es sich um eine "normale" Nummer, bei deren Anruf die auch im privaten alltäglichen Leben anfallenden, üblichen Telefongebühren entstehen.

* Das Informationssystem erwartet als sprachliche Eingabe die Angabe einer in Nordrhein-Westfalen vorhandenen Autobahnnummer. Dabei kann diese Eingabe in verschiedener Form erfolgen. Wünschen Sie beispielsweise Informationen über den Verkehrszustand auf der A-57, so kann dies entweder durch die sprachliche Eingabe als

- o "fünf sieben" oder als o "A fünf sieben" oder als
- o "fünfundsiebzig" oder als o "A fünfundsiebzig" erfolgen.

* Es ist nicht notwendig, Pausen zwischen den einzelnen Ziffern oder Wörtern einzufügen. Die Ziffer "2" kann sowohl als "zwei" als auch als "zwo" gesprochen werden.

* Neben der Spracherkennung kann das System auch mittels Tonwahl (DTMF) angesteuert werden. Durch Drücken der entsprechenden Folge von Zifferntasten kann bei Telefonen, die über das heutzutage übliche Leistungsmerkmal der Tonwahl verfügen, die gewünschte Autobahnnummer eingegeben werden. Innerhalb der Eingabe einer Autobahnnummer kann keine Kombination aus Sprach- und Tastatureingabe erfolgen (z.B. jemand spricht zunächst die Ziffer "fünf" und drückt anschließend die Taste "7"). Als Ersatz für das Kommandowort "ja" kann die Taste "*" und für "nein" die Taste "#" gedrückt werden.

* Beantworten Sie die Frage nach einer weiteren Stauauskunft mit "nein", so verabschiedet sich das System und trennt die Telefonverbindung.