en
cs de
sk es

TECH pedia

MMI (MULTIMODALE SCHNITTSTELLEN) FÜR DIE INTERAKTIVE KOMMUNIKATION UND NAVIGATION

JURAJ KAČUR, GREGOR ROZINAJ, RENATA RYBÁROVÁ

Titel der Arbeit: MMI (multimodale Schnittstellen) für die interaktive Kommunikation und Navigation

Author: Juraj Kačur, Gregor Rozinaj, Renata Rybárová

Übersetzt (von): Radoslav Vargic

Veröffentlicht (von): České vysoké učení technické v Praze

Fakulta elektrotechnická

Kontaktadresse: Technicka 2, Prague 6, Czech Republic

Tel.: +420 224352084

Drucken: (nur elektronisch)

Anzahl der Seiten: 40

Ausgabe: 1. Ausgabe, 2017

ISBN 978-80-01-06243-2

TechPedia
European Virtual Learning Platform for Electrical and Information Engineering

http://www.techpedia.eu

Erasmus+

Dieses Projekt wurde mit Unterstützung der Europäischen Kommission finanziert. Die Verantwortung für den Inhalt dieser Veröffentlichung (Mitteilung) trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

ERLÄUTERUNG

	Definition(en)
	Interessantheit (Interessantes)
	Bemerkung
	Beispiel
	Zusammenfassung
	Vorteile
	Nachteile

ZUSAMMENFASSUNG

Das Multimodal Interface (Eng Multimodal-Schnittstelle -. MMI) besteht aus mehreren Themen und Modulen, die für die natürliche und intuitive Kommunikation mit dem System verwendet werden. Alle zusammen stellen die MMI-Funktionalität dar. Zu den Modulen, die Teil der multimodalen Schnittstelle sein können, gehören: Sprechererkennung, Sprach- und Befehlserkennung, Gesichtserkennung, Gesten Navigation, Auge Navigation und Sprachsynthese.

ZIELE

Das Hauptziel des Moduls ist es, in die Grundlagen der multimodalen Schnittstellen und multimodalen Anwendungen einzuführen. Die Studierenden werden mit den Grundprinzipien der Gesichtserkennung, Sprechererkennung, Navigation über Sprachbefehle und Gesten vertraut gemacht. Am Modulende wird ein möglicher Einsatz der behandelten Module in Multimedia-Anwendungen beschrieben.

LITERATUR

[1] Probst, Michael; Schumann, Sebastian; Rozinaj, Gregor; Minarik, Ivan; Rybárová, Renata; Oravec, Miloš. EVALUATION: Final Multimodal Interface for User/Group-Aware Personalisation, Deliverable 5.5.1, available at http://www.hbb-next.eu/index.php/documents, Decmber 2013.

[2] Vančo, Marek; Minárik, Ivan; Rybárová, Renata. Evolution of static gesture recognition. In: Redžúr 2014 proceedings; 8th International Workshop on Multimedia and Signal Processing; 13 May 2014, Dubrovnik, Croatia. Bratislava: Nakladateľstvo STU, 2014, p. 41-44. ISBN 978-80-227-4162-0.

[3] Rautaray, Siddhart S., Agrawal, Anupam. Vision based hand gesture recognition for human computer interaction: a survey, Springer, 2012, pp 106-113. ISBN 978-3-642-27386-5

[4] Kristensson, Per Ola; Denby, L.C. Continuous recognition and visualization of pen strokes and touch-screen gestures. In Proceedings of the Eighth Eurographics Symposium on Sketch-Based Interfaces and Modeling, 2011, pp. 95-102.

[5] Gülbahar, Mark; Probst, Michael; Schumann, Sebastian; Rozinaj, Gregor; Minarik, Ivan; Rybárová, Renata; Oravec, Miloš. Design and Protocol: Intermediate Multimodal Interface and Context Aware Recommendation Engine, Deliverable 5.3.1, available at http://www.hbb-next.eu/index.php/documents, October 2012.

[6] T. Kinnunen, H. Li, An overview of text-independent speaker recognition: from features to supervectors, Speech communication, Vol. 52, pp. 12-40, Elsevier, 2010

[7] Benesty J., Sondhi M. M., Huang Y., Springer Handbook of Speech Processing, Springer-Verlag, ISBN 978-3-540-49125-5, 2008

Inhaltsverzeichnis

1 MMI Architektur

Aktuell sind die am häufigsten verwendeten Schnittstellen für Mensch-Computer-Kommunikation Tastatur, Maus oder Touchpad. Diese Geräte sind Schnittstellen zwischen Menschen und Computer. In den letzten Jahren wächst die Anforderung, dass Menschen mit Maschinen in der gleichen Art und Weise kommunizieren müssen, wie sie es untereinander machen: durch Sprache, Mimik und Gestik. So können die Menschen weit mehr Informationen übertragen als durch die Verwendung von Peripheriegeräten. Dies führt uns zu dem Begriff multimodale Schnittstelle (MMI).

Die Multimodal-Schnittstelle umfasst mehrere Bereiche und Module, die für eine natürliche und benutzerfreundliche Kommunikation mit dem System verwendet werden und bilden die MMI. Folgende Dienste können Teil der Multimodal Interfaces sein:

- Spracherkennung
- Spracherkennung und Sprachbefehle
- Multi-Gesichtserkennung
- Gesten Navigation
- Augen Navigation
- Sprachsynthese
- Hinweis Modul (Modul oder System das Informationen filtriert, um zu versuchen, die Bewertung oder Präferenzen vorherzusagen, die der Benutzer dem Element geben würde)

Die allgemeine Architektur der multimodalen Schnittstelle besteht aus mehreren Schichten. Die Bitübertragungsschicht stellt die Hardware-Eingabe- und Ausgabegeräte dar, die Interaktion mit der realen Welt ermöglicht. Multimodalen Daten, die von Eingabegeräten (Kamera, Sensor, ein Mikrofon, etc.) stammen, werden in weiteren Modulen parallel verarbeitet. Der MMI-Controller sammelt Ausgangsdaten von allen Modulen und bewertet den Ausgangsdatenstrom (Stream). Er enthält Informationen über die erkannten Benutzer und ihre angeforderten Aktionen.

2 Benutzeridentifikation

Heute werden Informationstechnologien zunehmend wichtiger. Sie begegnen uns in unserem täglichen Leben, beginnend mit Smartphones und endend mit anspruchsvolleren Geräten.

Jedes Gerät, das mit unseren persönlichen Daten arbeitet, sollte deshalb mit einer Art von Sicherheitssystem geschützt werden. Der am häufigsten eingesetzte Authentifizierungstyp ist ein Passwort-Schutz, aber mit der Entwicklung der Informationstechnologien und Sicherheitsschutzalgorithmen, Systemen und Anwendungen beginnt die Anwendung biometrischer Faktoren für die Authentifizierung. Dies hat mehrere Vorteile: Biometrie für eine Person ist einzigartig, so genügt die Anwesenheit einer Person vor dem Sensor und es besteht keine Notwendigkeit, sich mehrere Passwörter oder PINs zu merken. Es gibt Identifikationssysteme die auf verschiedene biometrische Merkmale wie Iris, Fingerabdruck, Stimme, Gesicht usw. auswerten. In diesem Teil werden wir uns auf die Gesichtserkennungs- und Sprecher- Identifikationssysteme im Allgemeinen konzentrieren.

2.1 2D- und 3D-Gesichtserkennung

Die Gesichtserkennung ist eine Teilmenge der Forschung auf dem großen Gebiet der Mustererkennung. Menschliche Gesichtserkennung ist in den letzten Jahrzehnten eine der wichtigsten Biometrie-Authentifizierungsmethoden geworden, vor allem wegen seines Potenzials für eine Vielzahl von Anwendungen und Bereichen (Überwachung, Sicherheit zu Hause, Grenzkontrolle usw.). Biometrische Systeme zur Personenidentifikation, die von mehreren Anbietern entwickelt werden, erreichen eine sehr hohe Gesichtserkennungsgenauigkeit. Die meisten dieser Anwendungen erfordern [1]:

- mehrere Gesichter von einem Video-Frame oder ein Bild erkennen zu können
- hohe Erkennungsrate
- Beleuchtungs-Invarianz
- Stabilität bei Gesichtsausdruckänderungen
- Erkennung in Echtzeit usw.

Der Hauptvorteil im Vergleich zu anderen Ansätzen ist, dass die Gesichtserkennung keine ausdrückliche Genehmigung des Benutzers erfordert, da Gesichtsbilder aus einer größeren Entfernung von einer Kamera erfasst werden können. Ein weiterer Vorteil ist, dass die Kosten für die Erfassungsgeräte sinken.

Anmerkung des Übersetzers: Rechtslage bezüglich Gesichtsbilderaufnahme kann in verschiedenen Ländern unterschiedlich sein.

Der Hauptvorteil der Gesichtserkennung ist die relative Einfachheit, die aber auch zu Ungenauigkeiten führen kann.

Im Vergleich zur Sprechererkennung, erzielt die Gesichtserkennung bessere Ergebnisse. Im Allgemeinen gibt es drei Hauptansätze basierend auf der Art der Daten, die in dem Erkennungsprozess verwendet werden. Hierzu gehören die Methoden für ein 2D-Intensitätsbild, 3D-Gesichtsdaten oder eine Kombination aus beiden. Der gesamte Prozess der Erkennung besteht aus drei Hauptstufen. Die erste ist die Akquisition und Vorverarbeitung, die zweite ist Datenregistrierung und die dritte Stufe ist die Erkennung. Detailbeschreibung zur 2D- und 3D-Gesichtserkennung kann im Modul zur *Benutzeridentifikation* gefunden werden.

In modernen Systemen kann Multi-Gesichtserkennung implementiert werden. In diesem Fall werden alle Gesichter in Bildern detektiert. Die Positionen der erkannten Gesichter werden mit dem Bild verbunden. Als nächstes wird das Bild in mehrere Teile (Proben) aufgeteilt, basierend auf den Positionen der detektierten Gesichter. Diese erzeugten Proben repräsentieren alle Gesichter in der Szene und erlauben separat jedes Gesicht zu verfolgen. Die Bildaufteilung ist ein Hauptschritt in Richtung Multi-Gesichtserkennung im System oder in der Applikation.

2.2 Sprecheridentifikation

Das allgemeine Ziel der Sprecheridentifikation ist es, automatisch zu entscheiden, wem die unbekannte Sprachprobe gehört. Die Entscheidung basiert auf einem Satz von Benutzersprachproben die in einer Datenbank während einer Trainingsphase gespeichert wurden. Wenn die beobachtete Entscheidungssicherheit zu niedrig ist, kann das System niemand identifizieren.

Mit einer begrenzten Anzahl von Benutzern für die Identifizierung, wird diese Art der Erkennung oft als geschlossene Gruppe Problem bezeichnet. Sprechererkennung wird seit vielen Jahren wissenschaftlich erforscht. Mit den neuen Technologien gibt es viele Anwendungsbereiche für die Identifikation, einige von ihnen sollen hier genannt werden:

- **Forensik**
- **Automatische Indizierung von Sprach- und Audioaufnahmen in Datenbanken**
- **Natürliche und nicht-invasive Methode für den Zugriff auf und sensible Daten, Dienstleistungen, usw.**
- **Anwendungen für die Spiele-Industrie**
- **Hilfsmittel für Menschen mit Behinderungen**

Sprechererkennung ist aus vielen Gründen ein komplexer Problemkreis. Die wichtigsten Herausforderungen sind:

- **Akustische Variabilität des Sprechers**
- **Änderungen verursacht durch aktuelle Gesundheit, physische und psychische Verfassung**
- **Vorhandensein von Hintergrundgeräuschen**
- **Einfluss eines Aufnahmegeräts und dem Ort der Aufnahme (Echos)**
- **Einfache Wiedergabe mit Aufnahmegeräten**

- Anderseits hat die Verwendung der Sprache als biometrisches Signal folgende Vorteile:
- **Keine Notwendigkeit Passwörter zu verwenden**
- **Nicht-invasive Verfahren**
- **Eine einfache Datenerfassung zur Herstellung eines biometrischen Signals**

Es gibt viele Anwendungen für Spracherkennung, die verschiedene Ebenen der Komplexität aufweisen wie: Anforderungen, Vertraulichkeit, Reaktionszeit, usw. So unterscheiden wir mehrere Hauptklassen von Identifikationssystemen:

- **Text unabhängige Systeme** – arbeiten unabhängig davon was gesagt wurde, so gibt es keine Anforderungen an die Phrasen, Sätze oder sogar Worte, außer vielleicht an die Länge einer echten Sprache, das heißt ohne Pausen. Der Genauigkeitsbereich ist dann irgendwo zwischen 80 bis 95%. Derzeit sind am häufigsten Systeme im Einsatz, die keine oder nur sehr wenigen Einschränkungen implizieren. Da sie niedrigere Genauigkeitsraten erreichen, können sie nicht in sicherheitskritischen Anwendungen eingesetzt werden.

- **Text abhängige Systeme** - erfordern präzise Sätze von einer unbekannten Person. Da solche Systeme Besonderheiten von Individuen nutzen können (so genannter Co-Artikulationseffekt), besitzen diese Systeme höhere Genauigkeitsraten von 95% bis sogar 99%. Außerdem können diese Systeme noch genauer und robuster sein, wenn der eingegebene Text ein Passwort spezifisch für jeden Benutzer ist. In bestimmten Fällen können diese Systeme den gewünschten Text durch verschiedene Fragen ändern,

- **Systeme basierend auf akustischen Informationen** - nutzen nur kurze akustische Informationen, die von der Sprache abgeleitet sind. Es ist vor allem eine Art des modifizierten und gefilterten Spektrums, das Formantenfrequenzen erfasst. Sie bilden die Grundlage für alle Systeme, da sie Informationen über die physikalischen Eigenschaften der Stimmorgane enthalten (Größen, Formen, Gewicht, Härte, etc.).

- **Systeme für Prosodie-Informationen** – werten die prosodischen Informationen aus, wie Rhythmus, Dynamik, Sprechen Tempo, Stimmung, Stimmmodulation usw. Diese Informationen sind aus längeren Zeiträume abgeleitet und sind schwieriger nachweisbar und quantifizierbar.

- **Systeme mit Merkmalen** höherer Ebene – benutzen die Informationen wie Dialekt, Sprache, Vokabular usw., aus welchen ein sozialer Status abgeleitet werden kann. Diese Informationen sind jedoch nicht diskriminierend genug.

- **Echtzeit-Systeme** - bieten sofortige Informationen über die Identität des Sprechers; eine gewisse Verzögerung von mehreren Sekunden ist für die Zuverlässigkeit erforderlich.

- **Nicht Echtzeit Systeme** – können komplexere Algorithmen nutzen um präzise Entscheidung zu liefern, eine ganze Sprachaufzeichnung wird verwendet.

Die oben genannten Aussagen und Ideen können sich über mehreren Bereichen der Wissenschaft verteilen. Für eine größere Einführung in die Sprechererkennung Problem lesen Sie bitte z.B. [6].

2.3 Mehrstufige Identifizierung

Für einige Anwendungen (z. B. App Store) können die Benutzer über eine oder mehreren Geräten-Arten zugreifen. Zu diesem Zweck wird eine mehrstufige Identifizierung durchgeführt.

Mehrstufige Identifizierung befasst sich mit verschiedenen Anforderungen für die Zugriffsrechte oder die Kontrolle über mehrere biometrische Modalitäten in Verbindung mit dem Namen, Passwort oder Pin. Aus der Gruppe der biometrischen Modalitäten wird für die Benutzeridentifizierung vor allem die Stimme und das Gesicht ausgewertet.

Je nach dem Ergebnis des Identifikationsprozesses wird der Zugriff auf das System oder die Anwendungen erlaubt oder kann eine weitere Überprüfung mit den Anmeldeinformationen - Benutzername, Passwort oder Pin erforderlich sein. In der Zukunft kann die Identifizierung um Iriserkennung erweitert werden und damit die Verwendung von PIN oder Passwort komplett ersetzen. Insgesamt wird damit die Sicherheit im System verbessert.

3 Kommunikation über Sprachbefehle

Die Kommunikation über Sprachbefehle fällt in das umfangreiche Gebiet der automatischen Spracherkennung (ASR).

Diese Bezeichnung bezieht sich auf automatische (Maschinen) Transkription der Rede. Der Eingang ist ein digitales Sprachsignal und der Ausgang ist eine Zeichenfolge, die die erkannten Wörter bezeichnet, die in dem verwendeten Wörterbuch enthaltenen sind. Da das gesamte Wörterbuch der Sprache sehr groß ist, in der Regel verwendet man nur eingeschränkte Versionen für einen bestimmten Einsatzbereich. Der Ausgabetext erfüllt die vorgeschriebene Grammatik oder beinhaltet die Phrasen die häufig in der vorhandenen Sprache verwendet werden.

ASR unterscheidet sich von der Aufgabe, den Sinn der Rede zu verstehen. Der Bereich, der diese Aufgaben behandelt wird künstliche Intelligenz genannt. Der Ausgangspunkt dafür ist ein Text, der nach der Anwendung von ASR entsteht.

In den letzten Jahrzehnten gab es eine intensive Forschung auf dem Gebiet der Konstruktion von ASR-Systemen, die in den folgenden Bereichen eingesetzt werden: Systeme der Bereitstellung von Informationen, Dialogsysteme, Hilfsmittel für Behinderte, Sprach-Transkription für die Archivierung usw. Es gibt die ersten nutzbaren Systeme, die außerhalb der kontrollierten Laborumgebung angewendet werden können, Deren Komplexität ist hoch, aber es gibt Anwendungen, die die kontinuierliche Sprache oder sogar ein normales Gespräch transkribieren können. Diese Aufgabe ist sehr komplex und muss die Probleme lösen, die in verschiedene wissenschaftliche Bereiche fallen. Praktische Systeme müssen in einer feindlichen Umgebung arbeiten (Änderungen von Hintergrund, Aufnahmegeräte usw.) müssen grammatische Vielfalt, akustische Variabilität des Sprechenden, große Wörterbücher usw. unterstützen. Ein typischer Benutzer fordert eine sofortige Antwort, ist nicht bereit seine Sprechart zu ändern, sich zu einem bestimmten Vokabular zu begrenzen und verliert schnell die Geduld mit möglichen ASR Systemfehler.

Es gibt eine große Anzahl von ASR-Anwendungen verschiedener Komplexität und mit unterschiedlichen Anforderungen und Randbedingungen. Dementsprechend gibt es mehrere unterschiedliche Klassifikationen derartiger Systeme nach verschiedenen Kriterien. Die bekannteste davon ist die Größe des Wörterbuches. Es werden folgenden Bereiche unterschieden:

- **Kleins Wörterbuch -** enthält bis zu hundert Worte
- **Mittleres Wörterbuch -** enthält hunderte oder tausende Wörter
- **Großes Wörterbuch -** bis zu hunderttausend Wörter

Diese Aufteilung ändert sich natürlich mit der Weiterentwicklung der Technologie.

Die Systeme können weiter sprecherabhängig bzw. sprecherunabhängig sein. Wenn ein System sprecherunabhängig ist, ist es egal wer spricht. Dies erfordert die Existenz eines allgemeinen Modells, das aus der großen Menge an Sprechenden abgeleitet wurde. Systeme die sprecherabhängig sind, sind für einen bestimmten Benutzer angepasst, und erzielen daher in der Regel bessere Ergebnisse, aber nur für den konkreten Benutzer. Systeme, die eine sofortige Reaktion liefern, werden Echtzeit-Systeme genannt. Eine bedeutende Rolle spielt dabei auch, in welcher Form die Spracheingabe erwartet wird. Dementsprechend sind die folgenden Szenarien zu unterscheiden:

- Isolierte Worterkennung - als Eingabe wird ein Wort aus dem Wörterbuch erwartet
- Diktier-Systeme – nehmen an, dass viele Wörter nacheinander folgen und es zwischen ihnen ausreichend gut erkennbare Pausen gibt.
- Kontinuierliche Rede – die Wörter fließen kontinuierlich, aber es gibt mehr oder weniger restriktive Regeln für Ihre Grammatik
- Natürliche Sprache - ist die schwierigste Aufgabe, behandelt die kontinuierliche Sprache ohne Einschränkung bezüglich Grammatik und in einer realen Betriebsumgebung

Eine Unterteilung ist auch nach dem modellierten Teil des Sprachmodells (Phoneme, Silben, Wörter, Phrasen, etc.) möglich bzw. wie sie es machen, z. B statistischer Ansatz mit *Hidden-Markov-Modells* **(HMM).**

Das Sprachsignal wird von einer menschlichen Sprachorgan erzeugt und wird als Vibrationsgeräusch wahrgenommen. Unter anderem trägt es die lexikalischen Informationen (was gesagt wurde). Die wird durch die entsprechende Sequenz von geeigneten Geräuschen repräsentiert, die spezifische akustische Form aufweisen. Diese akustischen Einheiten reichen aus, um eine bestimmte Sprache zu erzeugen und werden Phoneme genannt. Ihre Anzahl kann variieren im Bereich von 40 bis 60, abhängig von der Sprache. Ihre aktuelle akustische Form ändert sich von Sprecher zu Sprecher und ist deutlich von den umliegenden Lauten beeinflusst (so genannte Koartikulation).

Um den Teil in der Rede zu unterdrücken, der keine lexikalischen Informationen enthält, ist es notwendig ein Extraktionsverfahren für geeignete Merkmale zu verwenden. Die übrigen Informationen können die Worterkennungsrate verringern. Die Informationen über den Sprechenden (seine Stimmung, Gesundheit und aktuelle körperliche und emotionale Lage) sind ein Hindernis.

Es ist daher eine primäre Rolle der Systeme zur Merkmalextraktion, nur die bestmöglichen Merkmale auszuwählen, was praktisch nicht vollkommen gelingen kann. Das Sprachsignal enthält etwa 100 kb/s Daten, während der lexikalische Inhalt nur etwa 10b/s der Daten beträgt, weshalb Algorithmen zur Merkmalsextraktion als effiziente Informationskompressoren arbeiten.

Eine gute Wahl ist daher das menschliche Gehör zu simulieren, es mathematisch zu beschreiben, für die praktische Anwendung zu vereinfachen und für eine einfache Zusammenarbeit mit übergeordneten Spracherkennungssystemen anzupassen.

Es gibt viele erfolgreiche Methoden, die entweder das menschliche Gehör (kritische Zone, die Volumen-Kurve, etc.) oder den Prozess der Sprachproduktion (Alarm, Vokaltrakt, ...) simulieren. Der Vorteil der Methoden, das Gehör zu simulieren, liegt daran, dass sie günstige Mechanismen integrieren die die Rede zwischen anderen ungewünschten Signalen der realen Umgebung hervorheben. Sie unterdrücken das normale Hintergrundrauschen und die Verzerrung.

Während der Untersuchungen in den letzten Jahrzehnten wurde festgestellt, dass eine große lexikalische Information in der Hüllkurve von Signalspektrum verborgen wird. Damit kann man z. B. einzelne Phoneme Klassifizieren und unterscheiden. Genauer gesagt, diese Information ist in der Position, Form und Anzahl der dominanten Frequenzkomponenten verborgen. Um dies in Abb. 3.1 zu zeigen, sieht man das Spektrum des Lautes "e", Hüllkurve und die Position der sogenannten Formantfrequenzen (dominanten Frequenzen zu den Vokaltrakt Resonanzen im Zusammenhang). Abb. 3.2 zeigt den zeitlichen Verlauf des Lautes „e“.

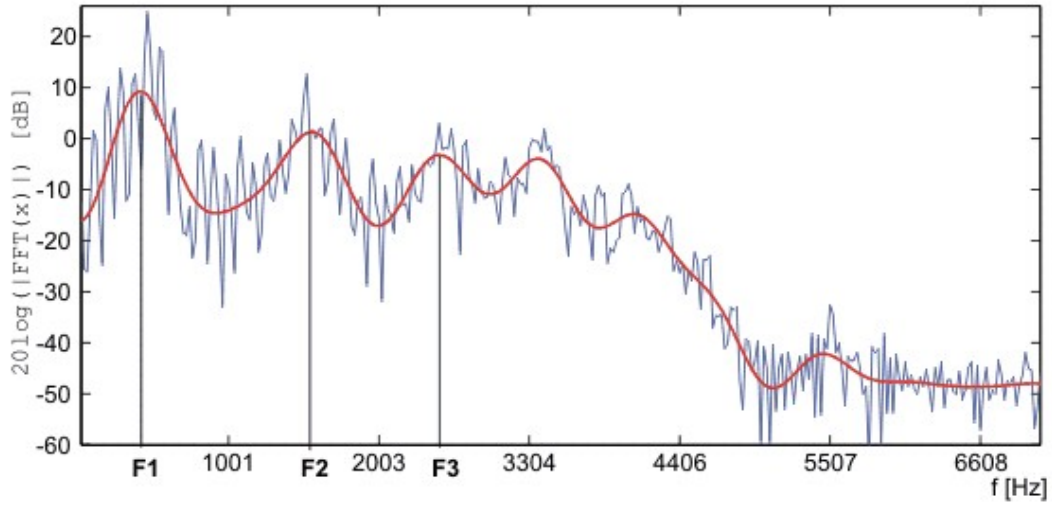

Abb. 3.1. Spektrum, Hüllkurve und Formantfrequenzen für den Laut „e“.

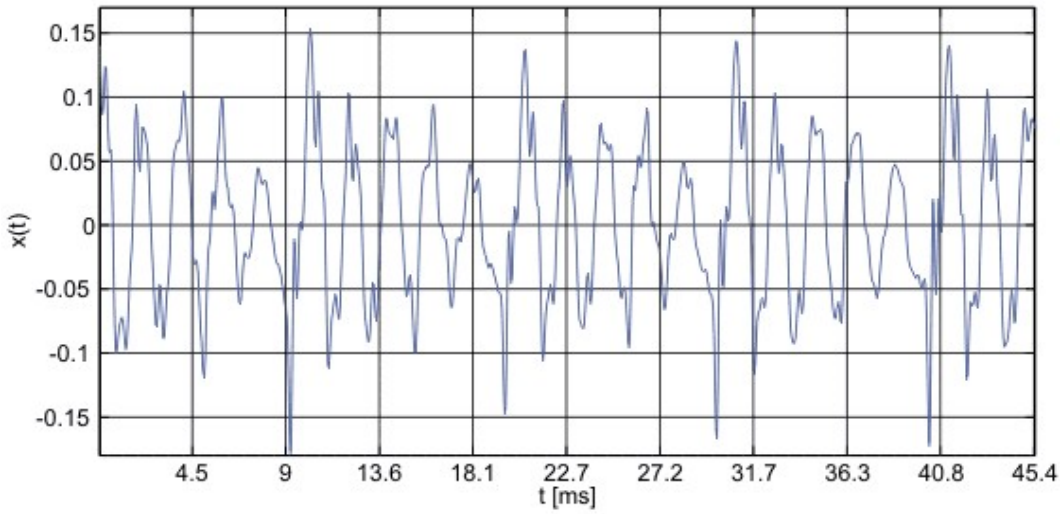

Abb. 3.2. Zeitlichen Verlauf des Lautes “e”.

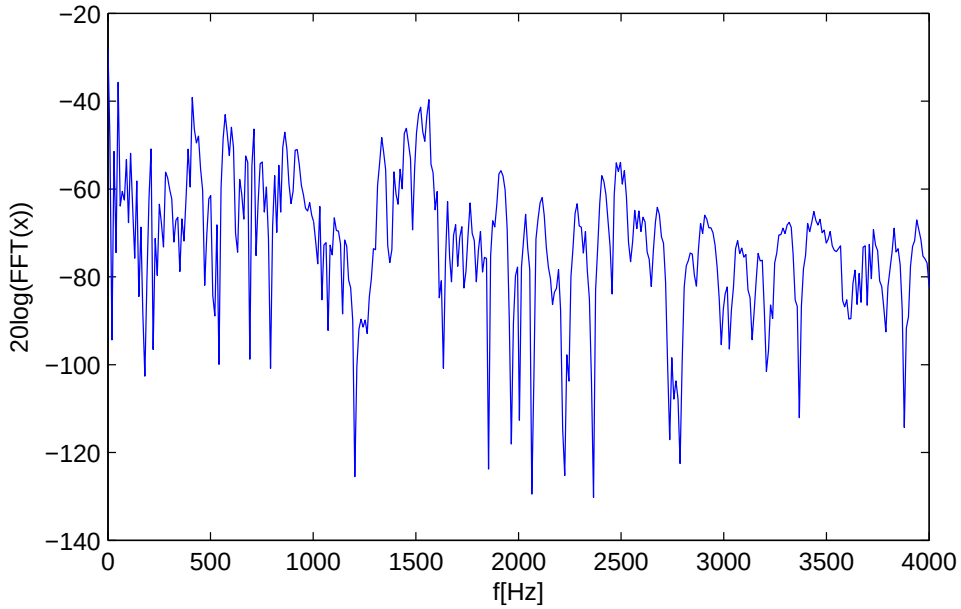

Abb. 3.3. Spektrum des Lautes “t”.

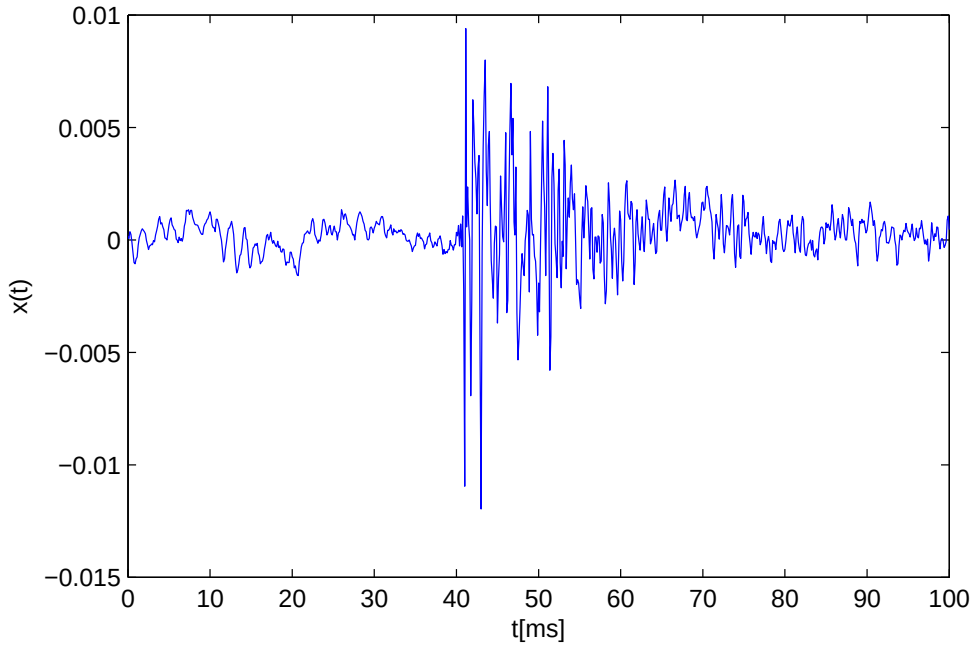

Abb. 3.4. Zeitlichen Verlauf des Lautes “t”.

Um den Unterschied zwischen Lauten zu demonstrieren gibt es in Abb. 3.3 Spektrum und in Abb. 3.4 den zeitlichen Verlauf des Lautes „t“.

In der folgenden Tabelle ist die Positionierung der ersten zwei Formantfrequenzen für Slowakische Vokale gezeigt und zwar getrennt für Männer und Frauen als Durchschnitt der Bevölkerung. Diese Tabelle enthält eine sehr grobe und einfache Art und Weise wie man die Konsonanten basierend auf ihrer Frequenz Form klassifizieren kann.

Tabelle 3.1. Positionierung der ersten zwei Formantfrequenzen für Slowakische Vokale, getrennt für Männer und Frauen

Vokal	Männer		Frauen	
	F1 [Hz]	F2 [Hz]	F1 [Hz]	F2 [Hz]
a	730	1100	850	1200
e	530	1850	600	2350
i	400	2000	430	2500
o	570	850	590	900
u	440	1000	470	1150

Das Merkmal-Extraktions-System soll in der Lage sein, die Unterschiede zwischen Formantfrequenzen zu benutzen. Auf der anderen Seite muss er "taub" sein für die Veränderungen, die natürlich und nicht hörbar sind. Die folgende Tabelle zeigt das hörbare und nicht hörbare Spektrum der Modifikationen.

Tabelle 3.2. Hörbare und nicht hörbare Spektrum der Modifikationen

Änderung typ	
Hörbar	**Nicht hörbar**
Anzahl von Formantfrequenzen	Die spektrale Neigung
Position von Formantfrequenzen	Frequenzen unterhalb von der ersten Formantfrequenz
Die Breite von Formantfrequenzen	Die Frequenzen oberhalb von der dritten Formantfrequenz
-	schmalbandigen Filter

Die Intensität des Geräuschs wird nicht linearen wahrgenommen. Dies kann durch die Logarithmus-Funktion approximiert werden.

In Bezug auf die Störung, sollten die Merkmale gegen additives Rauschen und auch Konvolution-Rauschen resistent sein.

Es muss möglich sein, die Merkmale leicht zu implementieren, mathematisch zu beschreiben, und sollte eine kompakte Form haben.

Es ist in der Regel gut, wenn die Merkmale untereinander linear unabhängig sind, um die Menge von Informationen zu verringern und anschließend die Modellierung zu erleichtern.

3.1 Parameter für die Spracherkennung

Es sollte beachtet werden, dass es die Parametrisierung einer Rede noch nicht gibt. Sie würde alle Eigenschaften für die optimale Parametrisierung erfüllen, die wir in im vorherigem Abschnitt erwähnt haben.

Daher ist die Forschung auf diesem Gebiet immer noch aktiv und es wurden verschiedene Arten gefunden. Die bekanntesten und besten sind Mel Frequency Cepstral Coefficients (MFCC) und Perceptual Linear Prediction (PLP). MFCC und PLP versuchen, das menschliche Gehör zu simulieren, das auf dem Gebiet der Spracherkennung zu guten Ergebnissen führt. Beide sind in der Lage die Lage, Anzahl und Breite der Formantenfrequenzen zu erfassen, die wahrnehmbar sind. Trotz der Ähnlichkeiten unterscheiden sie sich in den enthaltenen psychoakustischen Prinzipien.

MFCC filtriert das Sprachsignal über Hochpassfilter (Unterdrückungseffekt der Schallabstrahlung aus dem Mund), sie teilt die Sprache in Hamming Fenster und berechnet das Spektrum mit FFT. Anschließend wird das Spektrum nichtlinear modifiziert und zwar nach Mell-Skala (entsprechend der psychoakustischen Wahrnehmung den Frequenzbereichen von Menschen) die das Spektrum gleichmäßig auf Frequenzbänder mit einer Überlappung von 50% aufteilt. In Frequenzbändern wird die Leistung berechnet. Dies simuliert eine psychoakustische Filterbank, siehe Abb. 3.5. In der letzten Phase wird ein Logarithmus und eine diskrete Cosinus-Transformation (DCT) angewendet. Der Logarithmus modifiziert die Intensität des Schalls, wie es von einer Person wahrgenommen wird, und reduziert die Abhängigkeiten zwischen den DCT-Koeffizienten.

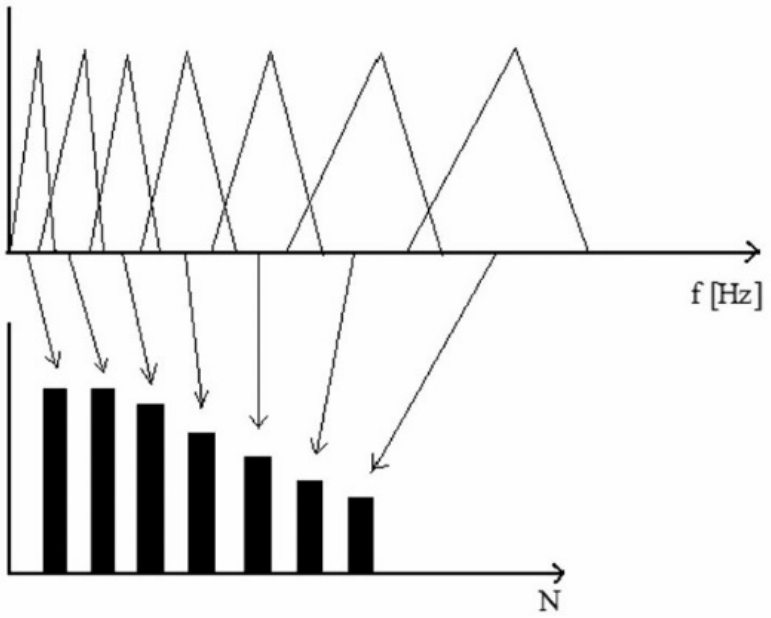

Abb. 3.5. Aufteilung des Spektrums nach Frequenzbänder in Mell Skala mit Filterbank und Dreieck Fenster.

PLP Symptome variieren in mehreren Aspekten: Sie verwenden die Bark Frequenzskala, glätten das Spektrum und Teilen ihn in Abständen von 1 Bark. Damit simulieren sie die kritischen Zonen für die menschliche Wahrnehmung (Definition), Gewichtsetzung für gleichmäßige Lautstarke, die Umwandlung von Energie nach Laustarke, Modellkoeffizienten Berechnung und ihre Transformation nach Kepstrum.

PLP ist komplexer als MFCC, enthält mehr psychoakustischen Prinzipien als MFCC, aber unter günstigen Laborbedingungen erreichen beide Methoden ähnliche Ergebnisse.

Dynamische Merkmale und Energie

Rede ist eine Folge von verschiedenen Klängen, so ist es zweckmäßig, den Fortschritt und die Veränderungen im Laufe der Zeit zu bewerten. Das gebräuchlichste Verfahren, das zu tun, ist die Anwendung von Differential- und Beschleunigungskoeffizienten, die aus Hauptmerkmalen berechnet werden.

Dies kann als eine einfache Differenz benachbarter Blöcke erfolgen, aber häufiger wird es als eine Linearkombination von Differenzen in einen größeren Zeitraum bestimmt. Es wurde gezeigt, dass die Energie die Position der Vokale gut lokalisieren kann . Dies erhöht die gesamte Diskriminierungskapazität der Merkmale. Daher wird oft die Energie (normalisiert) zu dem resultierenden Merkmalvektor addiert.

3.2 Spracherkennungsmethoden

Nach der Parametrisierung ist die Rede durch eine Folge von Merkmalvektoren gegeben, z. B. MFCC. Grob gesprochen, der Erkennungsprozess nimmt Muster oder Modelle der bekannten Sprache (Phoneme, Silben usw.) und vergleicht sie mit der unbekannten Rede. Dann werden die Proben oder Modelle mit der größten Übereinstimmung als erkannte Wörter erklärt.

Aufgrund der spezifischen Merkmale der Sprache (wie zum Beispiel Verschiedene Reden, die sich in der Länge sowie der Frequenz unterscheiden) ändert sich die Länge der Variation unregelmäßig innerhalb des Wortes. Bestimmte Teile könnten kürzer oder länger sein. Einfache lineare Dehnung (lineare Interpolation) oder Verkürzung bringt nicht den gewünschten Effekt. Die resultierende Rede wird als eine Kette von einzelnen Teilen (Modellen) der Sprache bestimmt, die einander folgen. Diese besonderen Eigenschaften führten zu Erkennungsmethoden (Mustererkennung), die für die Rede geeignet sind. Die bekanntesten von ihnen sind das *Dynamic Time Warping* (**DTW**) und die *Hidden-Markov-Modelle* (**HMM**). Natürlich gibt es viele Modifikationen, oder auch Kombinationen für verschiedene Anwendung sind geeignet. Im nächsten Abschnitt werden kurz diese zwei grundlegenden Methoden beschrieben, d.h. DTW und HMM.

DTW

DTW steht für die Dynamic Time Warping Methode, die akustisch zwei Merkmalvektoren mit verschiedener Länge vergleicht. Sie basiert auf nichtlinearem Zeitbiegeindex von Referenzsequenz und unbekannten Sequenz, bewahrt die Struktur der Sprache und versucht den kleinstmöglichen Abstand zwischen Sequenzen zu erzielen. Dies führt zu einer nichtlinearen Kompensation von Differenzen zwischen den Längen der beiden Sequenzen.

Die logische Bedingung ist, dass die Start- und Zielvektoren der zwei Sequenzen aneinander ausgerichtet sind. Dies erfordert die Verwendung des Algorithmus zur Erkennung von Anfang und Ende der Rede, die selbst eine relativ komplexe Aufgabe sein kann. Natürlich gibt es Änderungen, die diese strenge Bedingung nicht exakt erfordern.

Kurz gesagt, soll dieses Verfahren eine Zuordnung zwischen den zwei Sequenzen finden, so, dass jedem Vektor der ersten Sequenz einen Vektor aus der zweiten Sequenz zugewiesen werden kann. Dies führt zu dem Ergebnis, dass einige Vektoren aus einer Sequenz auf verschiedene Vektoren der zweiten Sequenz mehrmals abgebildet werden können. Das gilt auch umgekehrt. Natürlich hat dieser Prozess einige grundlegendende logische Einschränkungen, vor allem: Beginn und Ende der Vektoren einer Sequenz müssen dem Partner Vektor der zweiten Sequenz zugeordnet werden, die Biegefunktion darf nicht sinkend sein (die Reihenfolge der Vektoren bleibt unverändert) und es bestimmt die maximale Differenz in der Position, die noch überwunden werden kann.

Bei der Berechnung der DTW werden zwei Matrizen (Matrix der lokalen und globalen Abstände) verwendet. In der Matrix der lokalen Abstände befinden sich

akustische Entfernungen zwischen Merkmalvektoren von Referenzsignalen und von unbekannten Redesignalen. Die globale Matrix wird verwendet, um den akkumulierten minimalen Abstand zu berechnen und den Weg zwischen den zwei Sequenzen zu finden, also für jedes Element der globalen Matrix existiert ein Mindestabstand und einen Pfad zum Startpunkt der Matrix (initiale Vektoren von beiden Sequenzen) der sich unten links befindet.

Diese Situation ist in Abb. 3.6 erklärt. Natürlich gibt es festgelegte Regeln, wie man von einem Punkt der Matrix der globalen Abstände zu einem anderen Punkt kommen kann, zum Beispiel darf der Pfad nicht sinken. Wenn der Prozess das Ende (obere rechte Ecke) erreicht, endet der Vergleich und die kürzeste Distanz ist gefunden. Mit der Rückverfolgung kann man bestimmen wie der Minimalpfad aussieht. Dieser Vorgang wird für jedes Wort in dem Wörterbuch durchgeführt und das Wort, das den kleinsten Abstand zu dem unbekannten hat, erklärt die Rede als erkannt. Dieses Verfahren ist praktisch für isolierte Worterkennungsaufgaben am besten geeignet.

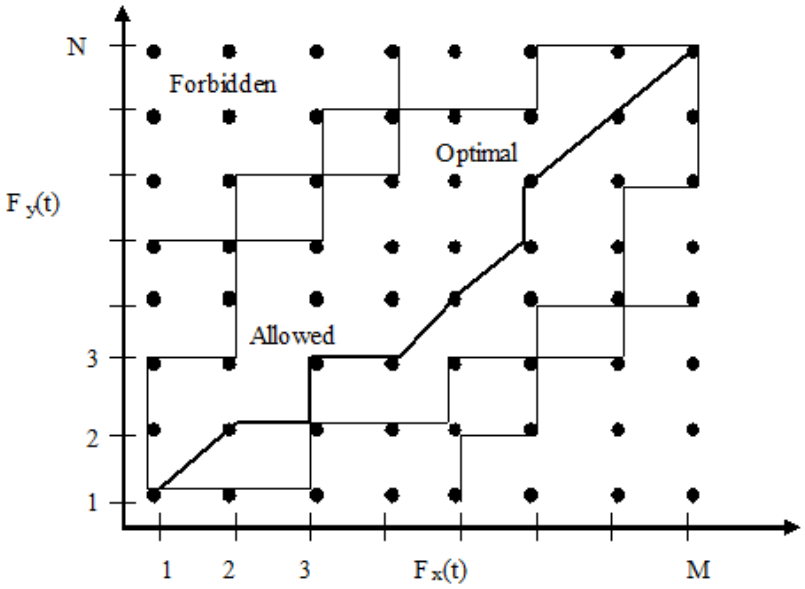

Abb. 3.6. Matrix der lokalen und globalen Abstände, Optimal Pfad und globale Einschränkungen.

DTW hatte große Bedeutung für die Spracherkennung vor allem für Probleme von isolierten Wörtern unabhängig von dem Sprechenden. Als sich die Anforderungen erhöhten (Sprecherunabhängigkeit, fließende Sprache) hat DTW seine wichtige Position verloren und wurde durch HMM-Methode ersetzt.

Hidden Markov Modelle (HMM)

Hidden-Markov-Modelle, gehören zu statistischen Modellierungstechniken die oft in der Spracherkennung verwendet werden. Diese Techniken erfüllen die Anforderungen an die Unabhängigkeit vom Sprechenden und eine einfache Verbindung der einzelnen Modelle (z. B. Phoneme) nach größeren Einheiten um

kontinuierliche Sprache zu beschreiben. Außerdem ist dieses Prinzip recht gut mathematisch beschreibend.

Für jede ausgewählte Spracheinheit (Phonem, Silbe, Wort, ...) wird ein HMM-Modell mit geeigneter Struktur geschaffen. Üblicherweise haben die Modelle der Sprachgrundteile wie Phoneme die gleiche Struktur und unterscheiden sich nur durch ihre freien Parametereinstellungen. Im Trainingsprozess werden diese freien Parameter so eingestellt, dass basierend auf einer Trainingssprach-Datenbank, alle statistisch signifikante Unterschiede in der Sprache entdeckt werden. Einzelne Reden werden so beschrieben, dass es offensichtlich ist, welche Folgen von Wörtern sie beinhalten. Am häufigsten sind die Modelle so eingepasst, dass die Trainings-Proben die höchste Wahrscheinlichkeit erreichen. Dieses Verfahren wird Training mit maximaler Wahrscheinlichkeit genannt. Es werden auch andere Kriterien verwendet die noch bessere Ergebnisse erzielen können, wie der größte Abstand zwischen den guten und schlechten Modellen, Fehlerminimierung (solche Verfahren werden Unterscheidungstraining genannt).

Jedes Modell besteht aus einer Anzahl von Zuständen, die ordnungsgemäß miteinander verbunden sind. Jeder Verbindung wird eine Übergangswahrscheinlichkeit (p) zugewiesen. Es gibt auch die Anfangswahrscheinlichkeit (π) die ausdrückt, dass sich das Modell gleich zu Beginn in diesem Zustand befindet. Dann kann ein solches Modell die Wahrscheinlichkeit des Auftretens einer Folge von Zuständen $S_1, S_2, S_3, \ldots S_N$ so beschreiben:

$$P(S_1, S_2, ..., S_{N-1}, S_N) = \pi(s_1)\, p_{12} \cdots p_{(N-1)N}$$

Weiterhin wird die zusätzliche Wahrscheinlichkeit zu jedem Zustand gegeben, welche eine Wahrscheinlichkeit des Auftretens des Merkmalvektor X im Zustand S, also $P(X/S)$ beschreibt. Damit ist die Verbindung zwischen den Zuständen und Vektoren der Sprachmerkmale bestimmt. Dann kann die Wahrscheinlichkeit der Beobachtungssequenz von Merkmalen $X_1, ..., X_M$ und Zustände $S_1, S_2, .. S_N$ einfach ausgedrückt werden als:

$$P(S_1, X_1, S_2, X_2, ..., S_N, X_N) = \pi_1 P(X_1 / S_1) p_{12} P(X_2 / S_2) p_{23} ... P(X_{N-1} / S_{N-1}) p_{N-1,N} P(X_N / S_N)$$

Beispiel des 4-Zustand Links-Rechts-HMM-Modell ist in Abb. 3.7 gezeigt.

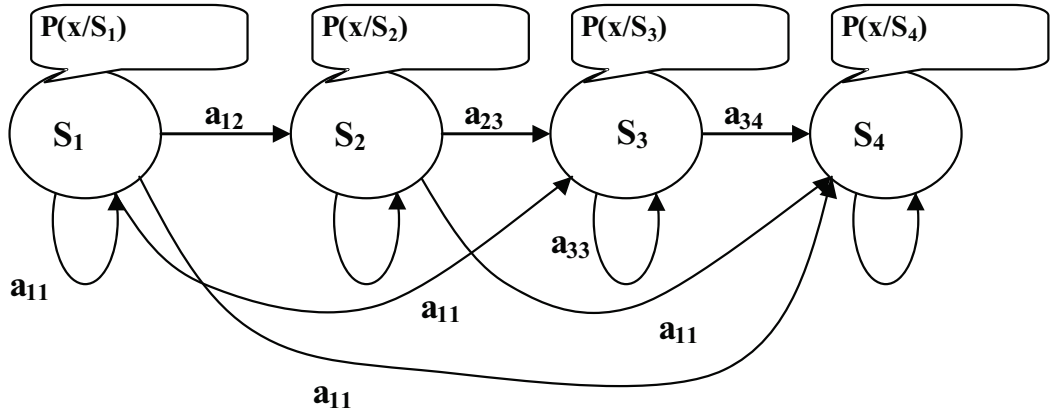

Abb. 3.7. 4-Zustand Links-Rechts-HMM-Modell.

Dann werden in dem Erkennungsprozess die Wahrscheinlichkeiten für die unbekannten Folgen von Merkmalen auf allen HMM Modellen in einem Wörterbuch (das Ergebnis der Konkatenation der Grund HMM-Modelle) berechnet und man wählt das Modell aus, das die höchste Wahrscheinlichkeit erreicht. Dieser Prozess ist schematisch in Abb. 3.8 dargestellt.

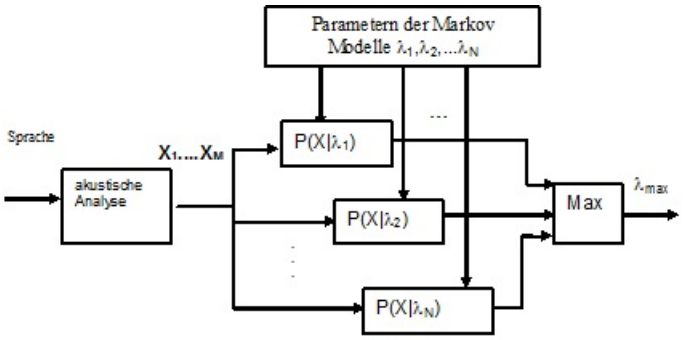

Abb. 3.8. Prozess der Spracherkennung mit HMM

Derzeitige fortgeschrittene Systeme, die den kleinsten *Wort Fehler* (**WER**) auf Basis von HMM erreichen, benutzen als Strategie für das Training: *maximale gegenseitige Information* (**MMI**), die *minimalen Klassifikationsfehler* (**MCE**) usw. Es werden auch Hybridverbindungen von HMM-Systemen und anderen Techniken wie Klassifikationssystemen insbesondere mit Support-Vektoren und tiefe neuronale Netze verwendet. Diese stellen nun die ultimative Spracherkennung dar und erreichen den niedrigsten WER für komplexe Aufgaben.

Nähere Informationen über die Spracherkennung und ihre Methoden können beispielsweise in [7] gefunden werden.

4 Navigation mit Gesten

Derzeit sind Gesten ein sehr beliebter Weg, um Anwendungen zu steuern, und die meisten Menschen nutzen sie jeden Tag. Wir benutzen Gesten in Mobiltelefonen, Computeranwendungen, Spielekonsolen usw. Es wird erwartet, dass in einigen Jahren das interagieren mit Gesten in fast jedem Gerät unterstützt wird. Die hohe Popularität der Navigation mit Gesten motiviert Forscher, diese Technologien zu verbessern. Navigation mit Gestern ist ein beliebter Trend und weil die Leistung des Computers kein Hindernis ist, erhalten sie viel Aufmerksamkeit [2].

Im Bereich der Gestenerkennung gibt es zwei Hauptbereiche, die sich entsprechend der Eingangshardware unterscheiden. Der erste Bereich ist für Zeigegeräte die sich auf die Geste konzentriert, wie Tablets, Touchpads oder Smartphones. Der zweite Bereich befasst sich mit Gesten die vor 2-dimensionalen oder stereoskopischen Kameras ausgeführt werden. Mit geringfügigen Modifikationen kann der gleiche Algorithmus zu Gestenerkennung für beide Bereiche angewendet werden. Darüber hinaus können die Algorithmen für stereoskopische Kameras modifiziert werden, um 3D Koordinaten für die Überwachung in allen drei Richtungen zu verwenden [2].

4.1 Klassifizierung von Gesten nach verschiedenen Kriterien

Gesten können abhängig von der Benutzererfahrung in zwei Kategorien eingeteilt werden, Angeborene Gesten, die auf die allgemeine Erfahrung aller Benutzer basieren, zum Beispiel das Objekt nach rechts mit einer Handbewegung nach rechts bewegen, das Objekt mit zusammengedrückten Finger ergreifen usw. Natürlich können angeborene Gesten durch Angewohnheiten und Kultur beeinflusst werden. Dies sollte berücksichtigt werden, wenn man Klassifizierung für bestimmte Gruppe von Menschen entwickelt. Bei Verwendung von diesen Gesten braucht man nicht die Gesten zu lernen, die Erfahrung ist ausreichend. Der Benutzer kann somit Anwendungen natürlich steuern. Die zweite Gruppe von Gesten sind die gelernten Gesten. In diesem Fall ist es notwendig, dass der Benutzer diese lernt. Gesten für Navigationssysteme und Anwendungen müssen einfach und natürlich sein und sollten wenig Kraft und Energie vom Menschen fordern.

Statische Gesten

Statische Gesten sind nicht von einer Bewegung abhängig. Statische Gesten stellen die Formen der Gliedmaßen dar, die Gesten darstellen und aussagekräftige Informationen tragen (Abb. 4.1).

Sie sind stationäre Art der Gesten, wo wir die Informationen über Bewegung nicht brauchen. Daher es ist nicht notwendig, die Reihenfolge der Bilder zu untersuchen, aber mindestens ein Bild. Die Sequenz mehrerer Bilder, die verschiedene Gesten enthalten, ist nur für ein höheres Maß an Verständnis wichtig.

Kontinuierliche Gesten

Kontinuierliche Gesten dienen als Grundlage für Interaktionen mit der Anwendung. Eine Kontinuierliche Geste stellt keine bestimmte Pose dar, wird aber verwendet, um mit der Anwendung oder in einer virtuellen Umgebung (Abb. 4.1) zu interagieren.

Ein typisches Beispiel für eine kontinuierliche Gestensteuerung ist eine Spielsteuerung durch kontaktlose Technologie, wie zum Beispiel bei Microsoft Kinect. Das System erfasst die Veränderungen in der Haltung oder Haltungsänderungen in einem Videospiel, es gibt aber keine spezifische Bewegung um eine Folge von Änderungen zu triggern.

Dynamische Gesten

Dynamische Gesten stellen Bewegung dar, die es Benutzern erlaubt, ein Objekt oder eine Anwendung direkt zu manipulieren (Abb. 4.1).

Dynamische Gesten können auf zwei Arten definiert werden. Erster Ansatz ist es, algorithmisch die Regeln und Bedingungen festzulegen, die die durchgeführte Geste erfüllen muss, um erfolgreich erkannt zu werden. Der zweite Weg sind

Templates. Templates sind Sätze von Punkten, die die Form der Geste identifizieren. Für die Bewertung der Erfolgsrate werden ausgefeilte Algorithmen verwendet, die Benutzerbewegungen mit den Templates vergleichen.

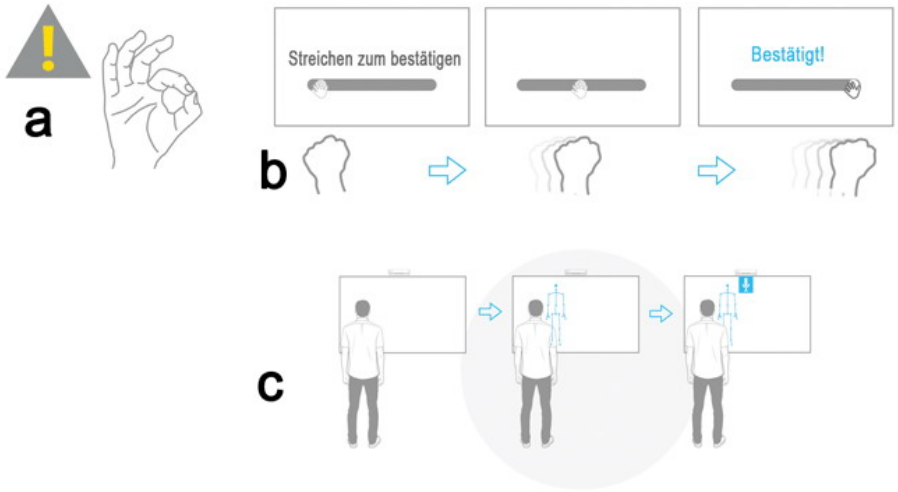

Abb. 4.1 – Arten von Gesten: a) statische Geste, b) dynamische Geste, c) kontinuierliche Geste

4.2 Bilddatenerfassung

Die Forschung für die Design-Gestenerkennungsalgorithmen begann mit der Auswertung von RGB-Werten (Rot Grün Blau) am Kameraeingang. Dieser Ansatz bringt einige Vorteile, aber die Nachteile überwiegen. Ein RGB-Bild besteht aus Farbkanälen und es gibt viele Faktoren, welche die Qualität von RGB-Bild negativ beeinflussen.

Die Beleuchtungs-Änderung kann die Bildqualität radikal ändern. Änderung von Lichtverhältnissen im einem Raum verändern die Helligkeit und den Kontrast des Bildes. Dies macht die Erkennung mit RGB-Bild unsicher und weniger präzis [2].

In den letzten Jahren begannen die Forscher Tiefenbilder für die Verarbeitung zu verwenden. Ein Tiefenbild ist ein zweidimensionales Bild, das zusätzliche Informationen enthält - die Tiefe. Die meisten Tiefensensoren basieren auf Infrarot (IR) Emission (d.h. der Kinect Sensor in Abb. 4.2). Der Sensor liefert die Daten die der Abstand jedes Pixels vom Sensor darstellt. Die erhaltenen Abstände können einfach nach Graustufendarstellung umgewandelt werden und die Tiefendaten werden als Graustufen-Videosequenz, dargestellt. Die Datenhaltung erfordert keine besonderen Bedingungen.

Die Tiefendaten haben gegenüber RGB-Kameras mehrere Vorteile. Der Hauptvorteil ist, dass die Tiefenkamera das Infrarot-Licht verwendet und im Infrarot-Bereich gibt es nicht so viele Störungen. Er ist nicht so empfindlich wie ein RGB-Sensor [2].

Abb. 4.2 – Das Bild aus Tiefenkamera

Der Abstand jedes Pixels des Sensors wird in Millimeter definiert.

Wenn wir Tiefenbild ins Graustufenbild konvertieren wollen, müssen wir die minimal und maximal mögliche Entfernung kennen.

$$Intensität = 255 - \frac{255 - (d - d_{\min})}{d_{\max}}$$

wobei d der Abstand des aktuellen Pixels ist, d_{min} der minimal mögliche Abstand von dem Sensor und d_{max} der maximal mögliche Abstand von dem Sensor [2].

Als Unterstützung für Eingangsbilderfassung können zusätzliche Gegenstände wie ein leuchtendes Armband, ein Ring oder ein kleiner Ball in der Hand dienen.

4.3 Methodik für Gestenentwurf

Die Problematik der Gestenerkennung ist nicht trivial. Beim Gestenentwurf müssen alle Einschränkungen und Unsicherheiten berücksichtigt werden, die bei der Eingabebilderfassung durch verschiedenen Geräte auftreten können. Die Methodik für den Gestenentwurf muss mehrere Anforderungen erfüllen: Natürlichkeit, Einprägsamkeit und einfachen Komfort. Komfort ist das wichtigste Kriterium für Anwendungen die durch Gesten gesteuert werden. Die nichtexistierende Standardisierung ist in jedem neuen Gebiet, also auch im Bereich der Gestensteuerung einer der Hauptprobleme für einen Anpassungsprozess. Wenn wir zufällig zehn Leute auswählen und sie bitten, eine Geste zu machen und die Musik startet, es ist sehr wahrscheinlich, das mindestens 5 verschiedenen Gesten vorgeschlagen werden. Das ist der Grund warum jeder Hersteller einen anderen Satz von Gesten verwendet. In der Praxis es ist schwierig , weil der Benutzer neue Gesten für jedes neue System lernen muss. Die Existenz von mindestens einem Standardsatz von Gesten könnte den Prozess der Gewöhnung auf Gestensteuerung in der Routinepraxis deutlich beschleunigen. Heute sind bereits Fernsehgeräte mit Gestensteuerung im Einsatz. Sie basiert auf der Bilderfassung mit einer 2D-Kamera. Sie ermöglichen eine einfache Menü-Steuerung des Fernsehers.

Trotz ihrer Mängel hat die Einführung von Gestensteuerung erzieherische Wirkung auf den Menschen, die sich allmählich an diese Art der Steuerung gewöhnen.

Die aktuellen Sätze von Gesten basieren auf Eingabegeräten (Tastatur, Maus, ...) die am Computer verwenden werden. Gesten versuchen im Grunde die vorhandenen Eingabegeräte zu ersetzen. Das Ziel des Gestenentwurfs war von Anfang an, unabhängig von den existierenden Benutzerschnittstellen, so nah wie möglich zu einem natürlichen freihändigen Betrieb zu gelangen, als ob die Gesten die einzige Möglichkeit zur Steuerung eines Gerätes wären.

4.4 Die Techniken für Erkennung der Gesten

Das übergeordnete Ziel der Handgestenerkennung ist die Interpretation der folgenden Möglichkeiten: die Hand Lage oder Haltung bzw. Geste.

Die Gestenerkennung kann auf zwei Arten durchgeführt werden. Entweder wird ein Datenhandschuh verwendet, der die Information über Flexion in eine elektronische Form verwandelt, oder es wird ein Vision-basierten Ansatz angewendet, wo eine Kamera als menschliches Auge dient und Körperpositionen aufzeichnet. Die Körperpositionen werden dann durch Bildverarbeitung extrahiert [5].

Es ist klar, dass die erste Methode präzisere Ergebnisse bringen kann, weil es sich um eine direkte und genaue Beobachtung des aktuellen Status des betreffenden Teils des menschlichen Körpers handelt.

Der Vision-basierte Ansatz, andererseits hat keine anderen Anforderungen an die Ausrüstung für den Endverbraucher (mit Ausnahme der Kamera) und ist für allgemeine Anwendung geeignet [5].

Ein Datenhandschuh ist ziemlich unangenehm in Bezug auf die Benutzerfreundlichkeit. Die benötigte Ausrüstung, um das Verfahren zu verwenden, ist für die meisten Standard-Kunden inakzeptabel teuer und ist nur für den speziellen Einsatz geeignet.

Der Nachteil des sichtbasiertes Verfahren liegt jedoch in algorithmische Komplexität weil eine beträchtliche Menge an Zeit und Rechenleistung erforderlich ist, um Körperbewegungen zu extrahieren [5].

Es stehen verschiedene Algorithmen zur Verfügung, welche sich auf die verschiedenen Aspekte und Annahmen der gestikulierten Person konzentrieren. Im Allgemeinen können die Algorithmen in zwei Kategorien unterteilt werden, und zwar Aussehen basierte Ansätze und 3D Modell basierte Ansätze. Die 3D-Modell-basierten Ansätze vergleichen die Eingangsparameter eines Gliedes mit einer 2D-Projektion seines 3D-Modells. Der auf Aussehen basierte Ansatz verwendet Bildmerkmale für die Modellierung der visuellen Erscheinung eines Gliedes und vergleicht sie mit Bildmerkmalen, die aus dem Videoeingang extrahiert wurden [5].

In Abschnitt 4.1 wurden drei Geste Typen definiert. Ein allgemeiner Klassifikator wird verwendet, um statische Gesten zu erkennen (das heißt Körperhaltungen). Unter dem Klassifikator verstehen wir - ein Merkmal oder eine Eigenschaft, basierend auf dem wir über die Zuordnung zu einer Gruppe oder Kategorie entscheiden. Der Klassifikator sagt aus, welche Geste oder Haltung erkannt wurde. Dynamische Handgesten haben auch einen zeitlichen Aspekt und erfordern Techniken, die diese Dimension behandeln, das heißt Hidden-Markov-Modelle (HMM). Ein anderer Weg ist die Verwendung von bewegungsbasierten Modellen.

Einige der Techniken für statische (und dynamische) Gestenerkennung sind K-means Algorithmen, KNN, SVM, der bereits erwähnten HMM-Algorithmus, der DTW-Algorithmus oder neuronale Netze [3].

Die Techniken für Erkennung der statischen Gesten

Für eine schnellere Datenverarbeitung arbeitet man im Prozess der Gestenerkennung nicht mit dem ganzen Bild.

Nachdem die Hand automatisch erkannt wurde, markiert man den Bereich um das Bild, nun wird nur dieser Bereich verarbeitet. Dies reduziert die Berechnungskomplexität. Erhaltene Abstände werden in Grautöne umgewandelt und so erhalten wir die Umrisse der Hand.

Im Folgenden Text werden zwei Verfahren beschrieben, welche für den Erkennungsprozess benutzt werden.

Methode der Konvex Mängel

Im ersten Schritt muss die Hand vom Hintergrundbild getrennt werden. Die Trennung kann man basierend auf dem Tiefenbild machen, wo bestimmt wird, ob das Pixel zur Hand gehört oder nicht. Der zweite Schritt ist die Kontur der Hand zu bestimmen.

Die Formen der zahlreichen komplexen Objekte sind gut durch konvexe Defekte charakterisiert. Abb. 4.3 zeigt das Konzept der konvexen Defekte auf dem Beispiel der Figur der Hand. Konvexe Hülle ist als eine dunkle Linie die den Handbereich markiert. Die Bereiche A bis H sind "Mängel" die zu diesem Paket gehören. Wie man sehen kann, charakterisieren diese konvexen Defekte den Zustand der Hand.

Der Algorithmus liefert die Koordinaten von drei Punkten, von dem Ausgangspunkt, tiefstem Punkt und dem Endpunkt (Abb. 4.4). Der tiefste Punkt kann man als der maximale Abstand zwischen der Hülle und den Konturen der Hand verstehen.

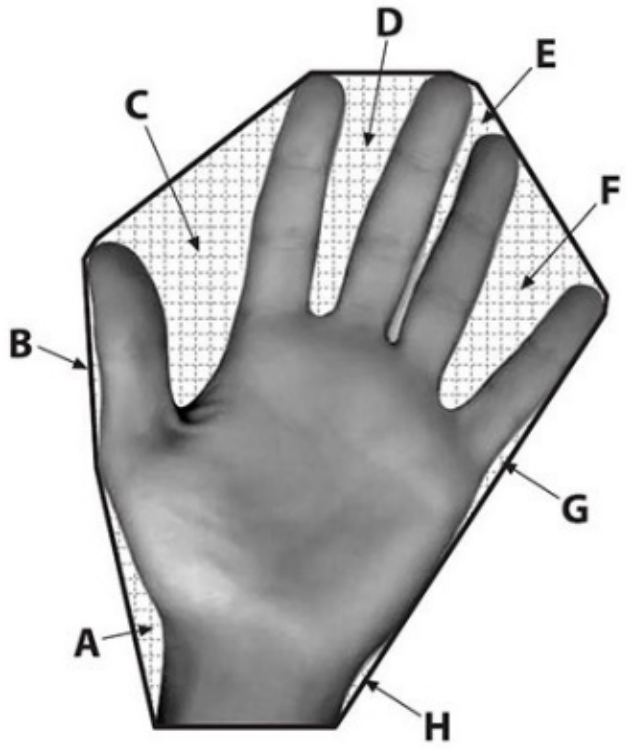

Abb. 4.3 –Konvexe Hülle dargestellt als dunkle Linie um die Hand

Abb. 4.4 - Der markierte Bereich stellt die konvexen Defekte dar und besteht aus drei Punkten: Ausgangspunkt (links), der tiefste Punkt (unten) und Endpunkt (oben)

Das Ziel des Algorithmus ist es, den Punkt auf dem Finger finden, der am weitesten von der Mitte der Handfläche entfernt ist. Der erste Schritt ist, alle Fehler zu beseitigen, deren Größe geringer ist als ein spezifizierter Wert. Dann entfernen wir die Fehler, die einen Abstand zwischen Ausgangspunkt und Maximalpunkt darstellen, der größer ist als ein bestimmter Wert und dynamisch definiert wird (wenn zwei Punkte zu weit sind, bedeutet dies, dass sie nicht einen Finger darstellen können). Als nächstes werden auch die Fehler entfernt, die einen geringeren Abstand zwischen Startpunkt und Tiefenpunkt besitzen als ein vorgegebener Wert. Dieser Wert ändert sich dynamisch entsprechend der Größe des Bereichs, wo die Hand erfasst wird (Abb. 4.5). Der letzte Schritt ist es, alle Defekte zu entfernen, die unterhalb des Handgelenks auftreten.

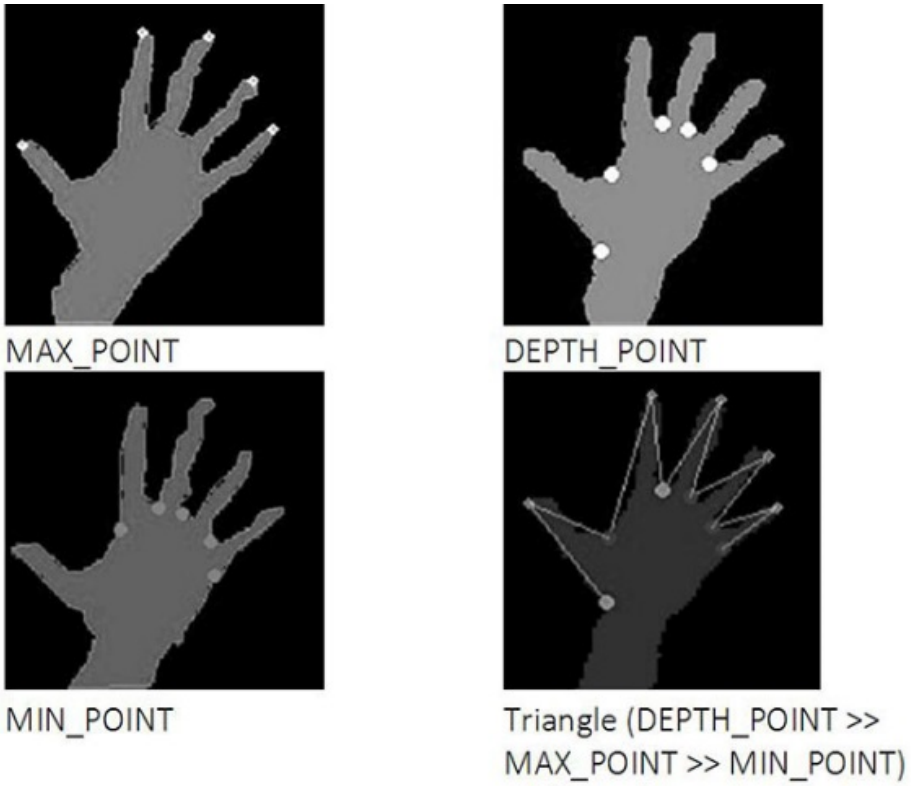

Abb. 4.5 – Darstellung der Logik die Dreiecke zum Minimieren der Fehler benutzt (das Bild oben links zeigt die maximalen Punkte, das Bild oben rechts zeigt die tiefsten Punkte, das Bild unten links zeigt die Minimalpunkte und das Bild unten rechts zeigt das resultierende Dreieck)

Erkennung von Handgesten basierend auf Teile

Im ersten Schritt des Algorithmus sind die Teile der Daten (Datenscheiben), die eine Hand beinhalten extrahiert und in ein binäres Bild umgewandelt. Die Mitte der Handfläche (Abb. 4.6) wird aus dem inneren Kreis und aus dem Hinzufügen eines Punktes auf der Hülle der Kontur berechnet, die Hand umfasst und hat einen maximalen Abstand vom entdeckten Fehler.

Abb. 4.6 – Die Suche nach Mitte der Handfläche

Alle Punkte der Kontur der Hand sind auf der X-Achse eingetragen. Die Y-Achse beschreibt den relativen Abstand der einzelnen Punkte von der Mitte der Handfläche. Alle eingetragenen Punkte bilden eine Kurve (Abb. 4.7).

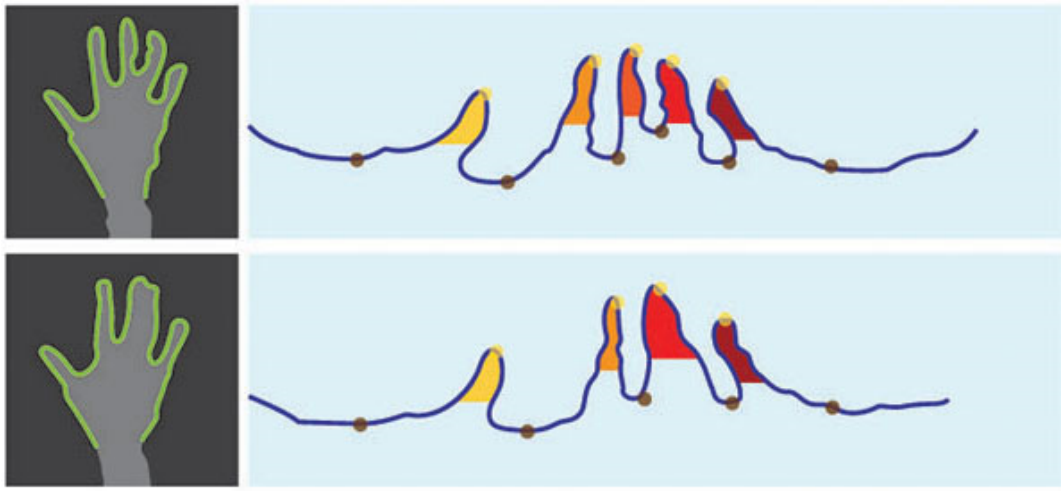

Abb. 4.7 – Kurvendarstellung

Der nächste Schritt ist die Kurve des Algorithmus zu analysieren und ein lokales Maximum zu finden. Nach einer vollständigen Analyse der Kurve und dem gefundenen Maximum, kann man mit der Fingersuche beginnen. Der Abstand von jedem Maximum wird mit dem eingestellten Grenzwert verglichen. Jeder Finger hat ein spezifisches Gewicht und Verhältnis zu den anderen Fingern. Darauf basierend können die Finger erkannt werden (das Gewicht wird mit vorgegebenen Werten verglichen und das ergibt die Anzahl von Fingern, d.h. wenn Gewicht <1,5 x das durchschnittliche Gewicht - Segment enthält einen Finger, usw.). Dieser Algorithmus funktioniert auch, wenn die Finger verbunden sind (Abb. 4.8).

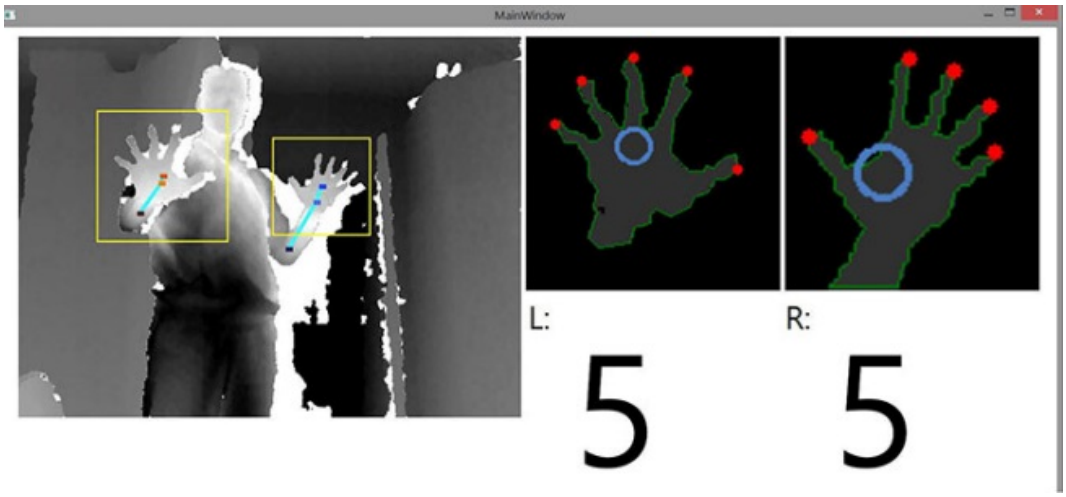

Abb. 4.8 – Erkannte Finger

Die Techniken für Erkennung der dynamischen Gesten

Dynamische Gesten erlauben einen Zugang zu geschützten Bereichen oder ermöglichen die Steuerung eines Systems oder einer Anwendung. Man kann sie auch als Passwortschlüssel verwenden. Die Gestenerkennung begann mit dem Einsatz neuronaler Netze und genetischer Algorithmen.

Diese Techniken hatten annehmbare Erkennungsrate, aber der größte Nachteil war die Menge der benötigten Rechenleistung und die benötigte Zeit, um die neuronalen Netze zu trainiert. Das war inakzeptabel für praktische Anwendungen.

Heutzutage werden neue Techniken der Gestenerkennung verwendet. Es wurden Algorithmen entwickelt, die neuronale Netze nicht erfordern, wie zum Beispiel Golden Section Search, Incremental Erkennungs-Algorithmus und Wahrscheinlichkeitsmodelle wie Hidden-Markov-Modell (HMM). Zur Erhöhung der Erfolgsrate der genannten Algorithmen kann man Maschinenlernen verwenden. In den letzten Jahren wurden die HMM Methoden sehr beliebt. Der Hauptgrund dafür ist, dass der HMM Ansatz wohlbekannt ist und in vielen Bereichen verwendet wird.

Der Algorithmus von Kristensson und Denby [4] ursprünglich vorgeschlagen für digitale Stiftstriche und Touch-Screen-Geräte kann auch für dynamische Gesten erweitert werden.

In dem Algorithmus wird eine Vorlage als eine Reihe von Segmenten definiert, um die Geste zu beschreiben. Es ist offensichtlich, dass eine Geste nur dann erfolgreich Erkannt werden kann, wenn eine genügende Anzahl von Segmenten ausgeführt wurden (Abb. 4.9).

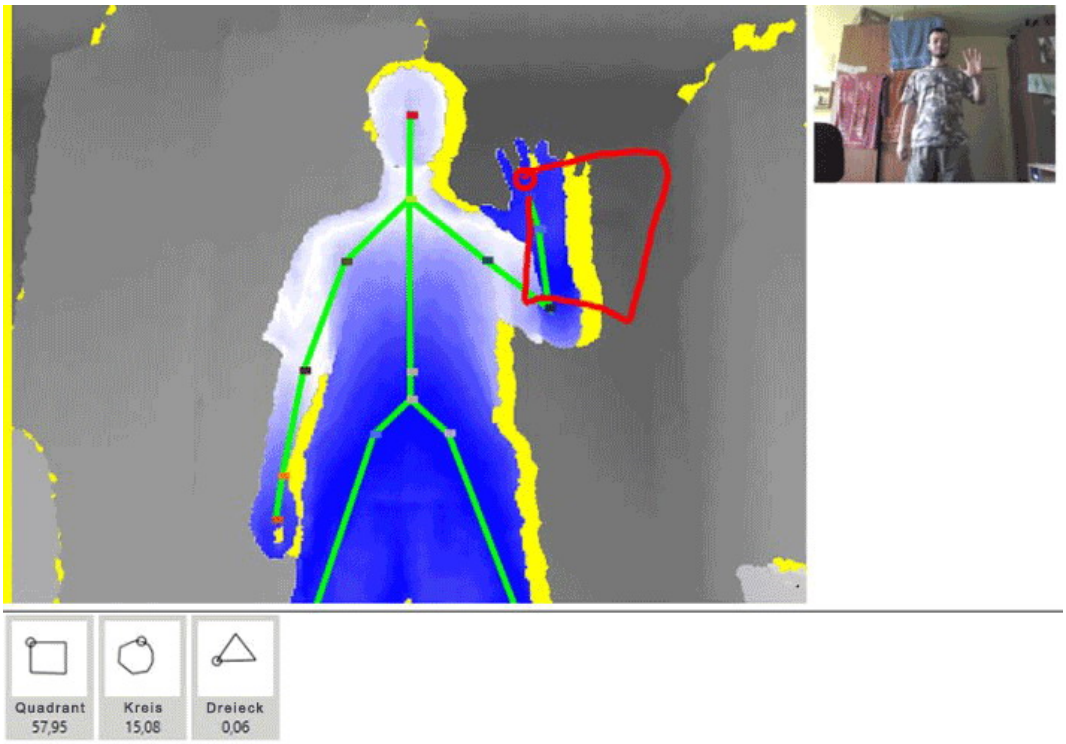

Abb. 4.9 – Vorschau der Testanwendung. Die rote Linie zeigt die ausgeführte Geste. Die Vorlagen unter dem Bild repräsentieren die wahrscheinlichsten Gesten (von links nach rechts). Für jede Figur kann man den Wert der Wahrscheinlichkeit für die entsprechende Geste sehen.

Die Verwendung von Gesten wurde um Streichgesten erweitert. Diese Art der Geste bringt einen sehr natürlichen und bequemen Ansatz. Streichgesten sind für schnelles navigieren in Menüs, Programmen und Galerien bestimmt und unterstützten vier Richtungen für jede Hand und ein paar Links-Rechts-Kombinationen. Die Methode Circle Dynamic Gesture Recognition (CNGR) wurde in [1] veröffentlicht und basiert auf Handerkennung, Bewegungsgeschwindigkeit und Abstand. Wenn die menschliche Hand schnellere Bewegung macht und den inneren Kreis verlässt verarbeitet das System diese Bewegung und bestimmt die Geste. Die Geste wird durch den Winkel der Bewegung vom mittleren bis zum äußeren Kreis gegeben. Diese Gesten sind möglich: Streichen nach links, rechts, oben und unten. Die gleichen Gesten können auch mit beiden Händen erfolgen. Der Benutzer kann auch eine Zoom-in und Zoom-out Geste durchführen (Abb. 4.10).

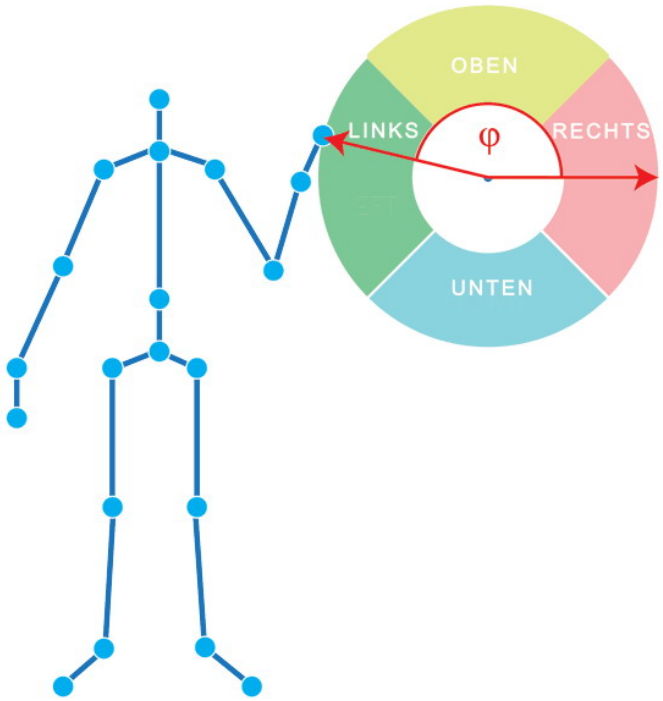

Abb. 4.10 – CDGR Gestenerkennung

4.5 Gestenerkennung in Praxis

Gestenerkennung kann als eine der ersten Möglichkeiten interpretiert werden, wie die Computer die Menschen oder die menschliche Körpersprache verstehen. Es kann zu dem großen Schritt führen, dass die Eingabegeräte wie Tastatur und Maus zurückgedrängt werden (auch in Sci-fi Filmen ist damit nicht gerechnet worden). Mit dem Konzept der Gestenerkennung ist es möglich, mit dem Finger auf dem Bildschirm zu zeigen, so dass sich der Cursor entsprechend bewegt.

Gestennavigation wird immer beliebter vor allem in Smartphones, Tablets und für moderne Fernseher. Jeder kennt Streichgeste für Bewegung nach oben/unten, nach links / rechts. Es gibt bereits viele Anwendungen die eine Reihe von Gesten zur Navigation in den Multi-Touch-Geräte anbieten, wie Lautstärkeregelung, den Bildschirm sperren, Steuerung von Musik-Player, Screenshot aufnehmen, navigieren zum Starpunkt, Aufruf neuer Anwendungen, Menü usw.

Das Anwendungsfeld für Gesten ist nicht beschränkt. Wir können Beispielsweise eine Tennistraining Anwendung mit 3D-Gestenerkennung von Cristian García Baeza und sein Team erwähnen.

Samsung hat 2013 zum Beispiel eine Gestenerkennung in ihrem "Flaggschiff" TV Modell vorgestellt. Die Erkennung besteht aus einem Tracking-Algorithmus mit Handgestenerkennung "Klick", die durch Schließen der Handfläche erzeugt wird. Wenn der Benutzer den Kanal wechseln will oder die Lautstärke einstellen will, muss er seine Hand heben und den Cursor mit der Hand so bewegen das er zu den Cursor-Tasten (analog einer Computermaus) kommt. Um die Lautstärke zu erhöhen, muss ein Benutzer "klickt" (das heißt schließen der Handfläche) so oft machen, bis die gewünschte Lautstärke erreicht wird. Dieser Ansatz könnte intuitiv zumindest für Computer-versierte Anwender betrachtet werden, aber kann auch als Beispiel für unnatürliche Gesten dienen.

5 GUI-Design

Die grafische Benutzeroberfläche (GUI) repräsentiert einen Weg für die Interaktion zwischen Mensch und Maschine. Die Benutzerschnittstelle umfasst Hardware (physischen) und Software (logischen) Komponenten.

Der Benutzer sendet Befehle oder Aufgaben an die Maschine/Anlage und die Maschine reagiert auf die gewünschten Befehle oder Aufgaben. In der Vergangenheit wurden GUI-Eingaben über Geräte wie eine Computertastatur und Maus realisiert. Als Touchscreen-Geräte erschienen, wurden GUI angepasst und auch die Eingabe durch Berührung von Fingern oder einem Stift wurde unterstützt.

Die Aufgabe, geeignete GUI für multimodale (oder jede andere) Anwendung zu erstellen, erfordern eine gute Praxis, die Erfahrung und das Feedback des Benutzers (heute wird dieses Forschungsgebiet User Experience genannt). GUI für multimodale Anwendungen sollten es auf einfache Weise möglich machen, eine Maschine effizient und benutzerfreundlich zu bedienen.

Im Idealfall braucht der Benutzer einen minimalen Aufwand für die Eingabe und die Maschine eine Mindestbetriebszeit um eine Aufgabe auszuführen. Das weiterführende Thema zur User Experience (UX) ist User Experience Design (User Experience Design UXD).

6 MMI Anwendungen

Moderne Anwendungsforschung ist auf natürliche multimodale Schnittstellen und deren Integration in das Multimedia-System für den Alltag fokussiert.

Die Verwendung von multimodalen Schnittstellen ist nicht nur auf das TV-System beschränkt, sondern hat viele andere Anwendungsgebiete.

Die Vision der zukünftigen Anwendungen und System ist nicht nur, den Fernseher und den Zugang zu Multimedia-Inhalte zu steuern, sondern ein System zu bauen, das Benutzer-Gewohnheiten und ihre Vorlieben und Interessen kennt. Ein solches System kann ein intelligenter Concierge des Hauses sein und dieser kann weiterte Verbindungen zwischen verschiedenen Dienstleistungen anbieten z.B. Einkaufen, während Sie, Voten oder ein digitaler Türsteher.

Für die bestmögliche Nutzung von multimodalen Schnittstellen, ist es wichtig zu wissen, welche Modalität mit welcher Aktion verbunden ist, und was auszuführen ist.

Für die Eingabe von Text wäre eine Eingabe durch Gesten schwierig, zeitaufwendig und unbequem, kann aber leichter und schneller mit Spracherkennung ausgeführt werden. Es ist also notwendig, eine Integrationsplattform einzuführen, die Daten für die Anwendungen liefern würden.

Einfaches Beispiel für ein mögliches Szenario: Die MMI-Steuerung für TV zeigt eine Liste der Benutzer, die wurden durch ein Sprach- oder Gesichtserkennungs-Module identifiziert. Nur Benutzern in dieser Liste wird erlaubt, TV-Anwendungen mit einem vordefiniertem Satz von Gesten, Sprachbefehlen oder andere Modalitäten zu steuern. Wenn der Benutzer den Raum verlässt, wird er automatisch aus der Liste entfernt.

Der andere Teil der Anwendung zeigt empfohlene Kanäle. Nach den Sehgewohnheiten des Benutzers, zeigt das System Empfehlungen, die am besten zu allen Benutzern vor dem TV passen.

Mit Streichgesten es ist möglich die Artikel in der Liste abzuspielen oder zu stoppen. Einige der empfohlenen Kanäle können gesperrt sein. Es bedeutet, dass sie diese Benutzer ohne Erlaubnis nicht abspielen können, bis sie das Geheimmuster eingeben. Geheimmuster werden durch dynamische Gesten ausgedrückt.