Automatische Spracherkennung

2 Grundlagen: die Vorverarbeitung des Sprachsignals

In diesem (längeren) Abschnitt werden grundlegende technische Begriffe aus dem Bereich der automatischen Spracherkennung (ASR) erläutert. In Zuge dieser Erläuterung bauen wir langsam eine Verarbeitungskette auf, die der Vorverarbeitung des Sprachsignals in einen modernen ASR-System entspricht.

Soweit vorhanden ist der jeweils englische Fachbegriff, so wie er in der Fachliteratur verwendet wird, in Klammern angegeben.

Beispiele von Computerkommandos und Skizzenanweisungen sind in `Schreibmaschinenschrift` wiedergegeben.

Literatur für diesen Abschnitt:
Reetz, Henning: Artikulatorische und akustische Phonetik, Wissenschaftlicher Verlag Trier, Trier, 2003, Kapitel 2.4 - 2.8.

Sprachsignal (speech signal)

Das Sprachsignal ist eine eindimensionale Funktion des Schalldrucks am Ort des Mikrophons über der Zeit (Zeitsignal). Es kann analog in Form einer elektrischen Spannung oder Stroms oder digital in Form von Werten (Zahlen) zu periodisch wiederkehrenden Zeitpunkten dargestellt werden.

Skizze

`Akustik` `Mikrophon` `Tiefpass` `A/D-Wandler` `Computer`

Analoge Sprachsignale können auf Tonbandgeräten aufgezeichnet und manipuliert (geschnitten) werden; digitale Sprachsignale können (da sie nur aus Zahlenwerten bestehen) wie andere Daten

auch in Computerdateien gespeichert und manipuliert werden.

```
Beispiel: Analoges Signal - Abtastung - Zahlentabelle
```

Die einzelnen Zahlenwerte sind im Prinzip Messwerte, die zu regelmäßig wiederkehrenden Zeitpunkten bestimmt werden. Sie werden auch Abtastwerte oder samples genannt. Entsprechend nennt man die Häufigkeit der Abtastung pro Sekunde die Abtastfrequenz oder Abtastrate (sampling frequency).

Praktische Übung: praat

```
Sprachsignal sei in dog.wav
- Praat aufrufen und dog.wav einlesen.
- Edit: in immer höhere Auflösung gehen, bis
  Signal 'eckig' wird.
```

Um ein analoges in ein digitales Signal umzuwandeln, muss das Shanon-Theorem oder die Nyquist-Bedingung erfüllt sein. Diese besagt, dass die Abtastrate mindestens doppelt so groß sein muss wie die höchste vorkommende Frequenz im abgetastetet Signal.

Nyquist/Shannon: $f_{abt} >= 2 f_{max}$

Diese Bedingung muss durch einen analogen Tiefpass vor dem A/D-Wandler gesichert sein. Wenn diese Bedingung verletzt wird (z.B. durch einen falsch gewählten Tiefpass oder eine zu niedrige Abtastrate), kommt es bei der Reproduktion des Signals zu Verzerrungen. Technisch gesehen 'spiegeln' sich hohe Frequenzen des abgetasteten Signals in tiefe Frequenzbereiche.

In der Praxis verwendet man heute fast immer Standard-Sound-Karten im PC als kombinierter Tiefpass und A/D-Wandler. Moderne Karten schalten automatisch entsprechend der gewählten Abtastrate einen passenden Tiefpass vor oder tasten zuerst mit sehr viel höherer Rate ab, filtern dann digital und reduzieren dann auf die gewählte Abtastrate.

Das Sprachsignal ist typischerweise im Bereich von 80 - 6000 Hz angesiedelt. Daher genügt für eine qualitativ hohe Aufnahme eine Abtastfrequenz von 16 kHz. Telefonkanäle werden nur mit 8 kHz abgetastet, was den typischen, etwas dumpfen 'Telefon-Sound' bewirkt.

Praktische Übung: esfilt oder genfilt

Tiefpassfilter 2000 Hz: Signal sei in dog.sfs

```
% esfilt -l -f 2000 dog.sfs
```

Telefonband: Signal sei in dog.sfs SP.01

```
% genfilt -i SP.01 -h 330 -l 3300 dog.sfs
% Es dog.sfs
% remove -i SP.02 dog.sfs
```

Die Rekonstruktion des digitalen Signals zum analogen Signal erfolgt ganz einfach durch das Verbinden der Abtastwerte in einem D/A-Wandler. Ein nachgeschalteter Tiefpass beseitigt die hochfrequenten Artefakte, die durch den 'eckigen' Kurvenverlauf entstehen. Das Ergebniss ist im Idealfall exakt das analoge Signal, das ursprünglich abgetastet wurde.

Digitale Sprachsignale kommen vor in:

- Telefonübertragung
- CD-Player (eine CD enthält durchschnittlich 325000000 Zahlen!)
- digitale Anrufbeantworter (oftmals ohne vernüftigen Tiefpass und daher mit recht merkwürdiger, 'spacy' Wiedergabe)
- DAT-Rekorder
- MP3-Player
- WAV-Dateien im Internet oder auf PC

Höhenanhebung, Preemphase (pre-emphasis)

Das Betragsspektrum (s. nächster Abschnitt) eines Sprachsignals fällt generell zu höheren Frequenzen mit ungefähr -6dB pro Oktave. (1 Oktave = Verdopplung der Frequenz; -6 dB entspricht einem linearen Faktor von 1/2). Dafür gibt es zwei Gründe: zum Einen bewirkt die Abstrahlung des Sprachsignals von den Lippen (Kugelwelle) einen Anstieg von +6dB pro Oktave, zum Anderen hat das Anregungssignal von der Glottis bereits ein fallendes Spektrum von ca. -12dB pro Oktave. Da sich Anregungsspektrum und Abstrahlungsfunktion multiplizieren (vgl. Quelle-Filter-Modell von Fant), addieren sich die Dezibel-Werte (logarithmische Skala): -12db + 6dB = -6dB
Praktisch bedeutet dies, dass das aus einem Sprachsignal berechnete Spektrum bei niedrigen Frequenzen höhere Werte hat als bei hohen Frequenzen. Dadurch werden Strukturen, insbesondere die Hüllkurve des Spektrums schlecht erfasst. Um diesem Effekt entgegenzuwirken filtert man *vor der Extraktion von Merkmalen* das Sprachsignal mit einem flachen Hochpass, der die hohen Frequenzen um +6 dB pro Oktave anhebt. Diese Filterung wird 'Höhenanhebung' oder 'Pre-Emphase' genannt.

Eine gängige Filterformel ist die einfache Differenzformel:

$s'(n) = s(n) - 0.95s(n-1)$

Skizze

`Skizze` `Normales Spektrum` `Filter` `Angehobenes Sprektrum`

Merkmale (features)

Als Merkmale eines Sprachsignals bezeichnet man zeitveränderliche Größen (Parameter), die eine Funktion des zugrundeliegenden Zeitsignals darstellen.

Es gibt mehrere Gründe für die Verarbeitung des Schalldrucksignals in Merkmale:

- hohe Variabilität des Sprachsignals
- Raumechos
- kategoriale Wahrnehmung von Merkmalen beim Menschen
- Physiologie: Vorverarbeitung im Innenohr
- Datenreduktion: Merkmale enthalten höhere Datendichte
- Robustheit: Merkmale sind robuster gegen Umwelteinflüsse und Störungen als das Sprachsignal

Die Berechnung der Merkmale aus dem Zeitsignal nennt man auch Merkmals-Extraktion (feature extraction).

Früher: Filterbänke und analoge (elektronische) Schaltungen;
Heute: fast ausschließlich Rechenvorschriften (Algorithmen).

Unterscheidung Kurzzeit- und Langzeit-Analyse, Fensterung

Skizze

```
                /   Fensterung, Analyse im Zeitraster
Zeitsignal
                \   Analyse ueber gesamte Aeusserung
```

Bei der Kurzzeitanalyse entsteht das Problem, dass kurze Signalstücke aus dem Sprachsignal 'herausgeschnitten' werden müssen. Ein solches Herausschneiden verursacht 'harte', d.h. nicht-stetige Übergänge an den Schnittstellen, was wiederum zu einer Verzerrung der analysierten Merkmale führt (besonders bei spektralen Merkmalen!). Um dies zu vermeiden, verwendet man sog. Fensterfunktionen (windows), mit denen das Sprachsignal an der Analysestelle multipliziert wird.

Bei der Kurzzeitanalyse wird über einen kleinen Teil des Sprachsignals (typischerweise 10 - 50 msec) ein sogenanntes Fenster (window) multipliziert, welches das übrige Signal ausblendet (vergleichbar einer Schablone, die man über ein Bild legt). Nur das 'noch sichtbare' Signal wird analysiert und in Merkmale verarbeitet. Die resultierenden Merkmale werden dem mittigem Zeitpunkt der Fensterlage zugeordnet. Das Ergebniss der Kurzzeitanalyse ist also eine Folge von Merkmalen (bzw. Merkmalsgruppen), die bestimmten Zeitpunkten zugeordnet werden können.

Skizze

```
Zeitsignal mit Hamming Window und
nächstem, überlappendem Window
```

Die Länge und Form des Fensters haben entscheidenden Einfluss auf die Ergebnisse der Analyse.

Beispiel:

```
Praat  - Sonagramm mit Breitband- und Schmalband-Filter
```

Es hat sich herausgestellt, dass für spektrale Analysen das sog. Hamming-Window am günstigsten ist. Es besteht aus einem leicht über die Nulllinie gehobenenem und abgeflachten Cosinus:

$s'(n) = s(n)[0.54 - 0.46cos(\frac{2\pi n}{N}]$

mit:

s'(n) : gefiltertes signal (n-tes sample im Fenster)
s(n) : ungefiltertes Signal
π : Kreiszahl 3.14...
N : Länge des Fensters in Samples

In der Spracherkennung werden typischerweise Fenster der Breite 20-25msec in einem Abstand von 10msec verwendet.

Bei der Langzeitanalyse wird das gesamte zur Verfügung stehende Signal zur Analyse berücksichtigt. Als Resultat ergibt sich ein Merkmal (bzw. eine Anzahl von Merkmalen) für die gesamte Äußerung. Meistens werden diese nicht direkt zur Spracherkennung verwendet, sondern dienen zur Normierung bzw. zur Adaption an langsam veränderliche Eigenschaften des Sprachsignals.

```
Beispiele: Mittlere Energie   -> Normierung
           Mittleres Spektrum -> Sprecheradaption
                                 Geräuschadaption
                                 Kanaladaption
           Mittlere Sprechgeschwindigkeit
                              -> Verschleifungen, Assimilationen
```

Merkmale werden zu sog. Merkmalsvektoren zusammen gefasst.

Einfache Merkmale (Kurzzeitanalyse) sind z.B.:

- Energie (Lautstärke)
- Pitch (Grundfrequenz)
- Formantlagen
- Energie in spektralen Bändern
- Lineare Prädiktion (LP)
- Nulldurchgangsrate

oder auch linguistisch definierte Merkmale:

- nasaliert
- stimmhaft/stimmlos
- hoch/tief
- vorne/hinter
- gerundet/ungerundet

Den Algorithmus zur Umwandlung des Sprachsignals in eine zeitliche Folge von Merkmalsvektoren nennt man auch 'front end'. Die Merkmalsvektoren werden in der Literatur häufig als 'frames' bezeichnet.

Die wichtigsten Merkmale in der ASR sind spektrale Parameter, sowie die Energie des Signals (Kurzzeitanalyse). Daher soll hier auf den Begriff des Spektrums allgemein und dann auf die zwei wichtigsten Formen von spektralen Parametern, dem Mel-Scale- oder Bark-Scale-Spektrum und die Lineare Prädiction (LP) kurz eingegangen werden.

Spektrum (spectrum, spectral coefficients)

Unter einem Spektrum versteht man zunächst ganz allgemein eine Zerlegung des Zeitsignals in sinoidale Anteile (Sinustöne verschiedener Frequenz, Phase und Amplitude), aus denen das Zeitsignal durch Superposition (= gewichtete Addition) synthetisiert werden kann. Eine solche Zerlegung nennt man Fourierreihenzerlegung oder Fourieranalyse oder auch Fouriertransformation. Der ungekehrte Weg, das Zusammensetzen eines Zeitsignals aus einzelnen Sinusanteilen nennt man auch Fouriersynthese.

`Praktische Übung:  harsyn`

Ein Spektrum kann man sich als dichte Aneinanderreihung von vielen Linien vorstellen, wobei jede dieser Linien den Anteil eines Sinustons mit einer bestimmten festgelegten Frequenz repräsentiert. Folgerichtig stellt man daher ein Spektrum über der Frequenz und nicht über der Zeit dar.

Die Berechnung des Spektrums - genauer des Kurzzeitspektrums, wenn es sich um ein kurzes Signalstück handelt, das mit einer Fensterfunktion aus einem längeren Signal herausmultipliziert wurde - wird durch die Anwendung der Fourierreihenentwicklung berechnet. In der Praxis verwendet man fast immer dafür die sog. Fast-Fourier-Transform (FFT), welche im Gegensatz zur normalen FT weniger Rechenaufwand erfordert.

Das resultierende Spektrum ist komplex und achsensymmetrisch zur höchsten analysierbaren Frequenz (siehe auch Shanon-Theorem). Daher interessiert primär nur die erste Hälfte des berechneten Spektrums.

Skizze

`Skizze  Zeitsignalstueck  Spektrum symmetrisch`

Aus dem Spektrum kann man sofort mehrere Eigenschaften des analysierten Sprachsignals erkennen:

- handelt es sich um ein Linienspektrum (einzelne Linien auf geradzahligen Vielfachen der Grundfrequenz), ist das Signal periodisch (stimmhaft)
- ist das Spektrum glatt, handelt es sich um ein impulsartiges Signal (stimmlos)
- ist das Spektrum regellos (verrauscht), so handelt es sich um ein statistisches Signal (Rauschen, stimmlos)

- Formanten (= Resonanzfrequenzen des Vokaltrakts) bilden sich als deutliche Gipfel im Spektrum aus.

Die Breite des Analyse-Windows bestimmt auch die Auflösung des berechneten Spektrums: Je schmaler das Window, desto schlechter die spektrale Auflösung; umgekehrt bedeutet aber ein breiteres Window eine schlechtere zeitliche Auflösung. Durch eine Aneinanderreihung von mehreren Spektren über einer Zeitskala erhält man ein Sonagramm (meistens wird dabei die Intensität der Frequenzanteile durch Schwärzung ausgedrückt).

```
        Praktische Übung:        anspect

        Sprachsignal sei in dog.sfs
% Es dog.sfs
  Zeitpunkt bestimmen, an dem das Spektrum berechnet werden soll: T
% anspect -i SP.01 -t T dog.sfs
Varianten mit verschiedener Fensterbreite:
10 msec
% anspect -i SP.01 -t T -w 0.01 dog.sfs
50 msec
% anspect -i SP.01 -t T -w 0.05 dog.sfs
```

Mel-Scale bzw. Bark-Spektrum

Ein Fourierspektrum enthält immer noch genauso viele Samples wie das zugrundeliegende Sprachsignal im Window, bzw. um genau zu sein, enthält es halb so viele Samples, weil man meistens nur die Hälfte des symmetrischen Spektrums betrachtet. Um zu robusteren spektralen Merkmalen zu kommen, ist es sinnvoll, Bereiche im Spektrum, sog. Bänder, zu integrieren und nur noch die Energie oder eine äquivalente Größe dieses Bandes weiter zu verarbeiten. Dadurch reduziert sich die Anzahl der Parameter pro Kurzzeitanalyse von z.B. 256 auf 12, also eine drastische Datenreduktion.

```
Beispiel Datenreduktion
Sprachsignal 10sec, 16kHz = 160000 Abtastwerte
12 Merkmale alle 10msec = 10 * 100 * 12 = 12000 Merkmale
     -> Reduktionsfaktor 160000 / 12000 = 13,33
```

Es hat sich gezeigt, dass Sprache am besten durch Bänder repräsentiert wird, welche die Frequenz-Orts-Transformation des menschlichen Innenohrs nachbilden.

In der Cochlea werden niedrige Frequenzen deutlich besser aufgelöst als sehr hohe Frequenzen; der Zusammenhang ist nicht- linear und wird in der Psychoakustik mit Hilfe der Mel-Scale abgebildet. Daher verwenden manche ASR-Systeme Merkmale, die gleich große Bereiche auf der Mel-Scale abdecken. Das bedeutet: der niedrige Frequenzbereich der Sprache (100-800 Hz) wird durch genauso viele Bänder repräsentiert, wie der physikalisch sehr viel breitere Bereich von 800-3000 Hz.

Praktisch erreicht man dies beispielsweise durch eine Integration des Spektrums in Bändern, die mit der Frequenz immer breiter werden; ein solches Spektrum nennt man 'gehörgerecht' oder ganz einfach Mel-Spektrum.

Näherungsformel für die Beziehung von mel zu Hz:

$mel(f) = 2595log(1 + f/700)$

f : Frequenz in Hz

Darüberhinaus lässt sich durch psychoakustische Versuche nachweisen, dass der Mensch die Lautstärke eines Signals nicht einfach durch seine Gesamtenergie bestimmt, sondern dass der Eindruck der Lautheit sehr stark durch die Form des Spektrums beeinflusst wird: Geräusche mit schmalem Spektrum (z.B. ein Flötenton) erscheinen sehr viel weniger laut als ein breitbandiges Geräusch mit der gleichen Energie (z.B. ein landendes Flugzeug). Die Bestimmung der Lautheit erfolgt beim Menschen durch die Integration der Energie in bestimmten Bändern, den sog. Frequenzgruppen (critical bands), die danach zu einer Gesamtlautheit zusammengefasst werden. Die Breite dieser Frequenzgruppen ist messbar und steigt analog zur Mel-Scale von tiefen zu hohen Frequenzen nicht-linear an. Die Breite einer Frequenzgruppe ist als die psychoakustische Einheit 'Bark' definiert, und das menschliche Gehör erstreckt sich im besten Falle (Kleinkind) über 24 Bark. Motiviert durch diese Tatsache verwenden ASR-Systeme auch häufig 20 - 22 Bark-breite Bänder zur Zusammenfassung des Spektrums zu Merkmalen.

Skizze

```
Spektrum   Zusammenfassung zu Frequenzgruppen 1 Bark = 100 mel
```

Lineare Prädiktion (linear predictor, LP)

Die lineare Prädiktion geht vom Quelle-Filter Modell Fants aus, welches besagt, dass jedes Sprachsignal vereinfacht gesehen aus zwei wesentlichen Anteilen besteht: dem Anregungssignal der Glottis und der spektralen Verformung dieses Signals durch den nachgeschalteten Vokaltrakt, der wie ein akustisches Filter wirkt.

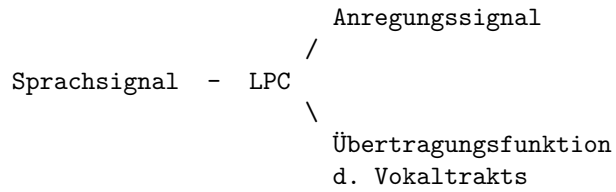
Skizze

Skizze	Quelle Anregungs-spektrum	Filter Uebertragungs-funktion	Sprachsignal Spektrum

Mathematisch gesehen ließe sich nach diesem Modell jedes Sprachsignal als die Faltung des Anregungssignals in der Glottis (die Stimmlippenschwingung) mit der Impulsantwort des Vokaltrakts berechnen und auch wieder analysieren. Im Spektralbereich entspricht einer Faltung im Zeitbereich ein einfaches Multiplizieren der beiden Spektren bzw. Übertragungsfunktionen (= Spektrum der Impulsantwort). Dies klingt komplizierter als es ist; praktisch müssen sich zum Beispiel in jedem stimmhaften Sprachsignal zwei spektrale Anteile feststellen lassen, wobei der eine Anteil das Spektrum der periodischen Stimmbandschwingung in der Glottis ist, und der andere Anteil die sog. Übertragungsfunktion ('Filterkurve') des Vokaltraktes, welche beide miteinander multipliziert das Spektrum des resultierenden Sprachsignals ergeben. (vgl. Seminar P4.1 'Einführung in die Signalverarbeitung')

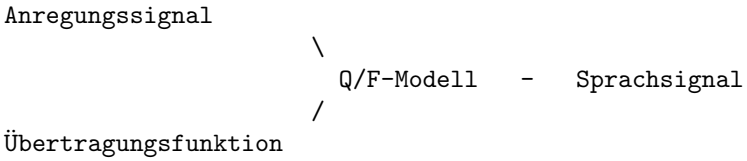
Die LP macht sich diese Modellvorstellung (die im übrigen auch für nicht-stimmhafte Signale angewandt werden kann) zunutze und versucht, durch eine inverse Filterung das Sprachsignal wieder auf die Anregung vor dem Vokaltraktfilter zurückzurechnen.

Skizze



Aus den beiden so ermittelten Spektralanteilen lässt sich unter Anwendung des Fantschen Quelle-Filter-Modells das Originalsignal wieder rekonstruieren.

Skizze



Für die Spracherkennung ist die Rekonstruktion uninteressant. (Für andere Techniken, z.B. die Übertragung von Sprachsignalen zwischen Handys ist dies sogar sehr interessant!) Das Anregungssignal wird verworfen und nur die Faktoren der Übertragungsfunktion des Vokaltrakts weiterverwendet, weil diese die wesentlichen phonetischen Informationen über die gesprochenen Worte (Lexik) enthalten.

Berechnung der LP Koeffizienten:

Aus den N zurückliegenden Abstastwerten wird mit Hilfe linearer Kombination der nächstfolgende Abstastwert $s'(n)$ vorhergesagt (daher 'Prädiktor').

$s'(n) = c_0 + c_1 s(n-1) + c_2 s(n-2) + ... + c_N s(n-N)$

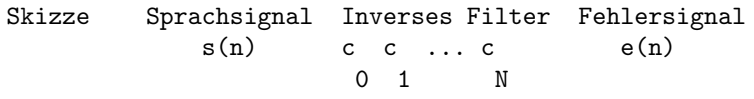
Durch Differenzbildung des vorhergesagten Wertes mit dem tatsächlich gemessenen Wert kann ein Fehlersignal $e(n)$ bestimmt werden, das in einem Optimierungsverfahren (für das Analysefenster) minimiert wird. D.h. die Koeffizienten des linearen Prädiktors werden so eingestellt, dass im Analysefenster im Mittel ein minimales Fehlersignal entsteht.

$e(n) = s(n) - s'(n)$

Optimierung der $c_0, c_1, ...c_N : e(n) -> 0$ für alle n

Die lineare Kombination des nächsten Abtastwertes aus seinen Vorgängern ist technisch nichts anderes als ein digitales Filter (genauer ein FIR-Filter), dessen Filterfunktion sich aus den Koeffizienten berechnen laesst. Diese Filterfunktion modelliert dann (im Idealfall) den Vokaltraktteil, während das nach der Filterung verbleibende Fehlersignal die Anregung durch die Glottis darstellt (man spricht hier auch von 'inverser Filterung'). Je nachdem, wieviele Abtastwerte N in die Berechnung einbezogen werden, erhält man logischerweise auch N Filterkoeffizienten. Man spricht dann von einer LP N-ter Ordnung.

Skizze



Die Filterkoeffizienten und damit das Vokaltraktfilter werden - analog zur Kurzzeitanalyse - für jedes Analysefenster berechnet und man erhält (zusätzlich zum Fehlersignal, das aber nicht weiter betrachtet wird) einen Koeffizientensatz, der die Form des Vokaltraktes zu diesem Zeitpunkt modellieren soll. Diese können nicht direkt als Merkmale verwendet werden, weil sie nicht isomorph sind. D.h. eine Übertragungsfunktion kann durch beliebig viele verschiedene Kombinationen von Filterkoeffizienten modelliert werden. Daraus folgt logischerweise, das man diese Filterkoeffizienten nicht als Merkmale für die Mustererkennung verwenden kann (in der Praxis wird es aber dennoch gemacht!). Aber die daraus resultieren spektralen Werte der Filterfunktion (häufig in Form von sog. 'Cepstren') stellen genau wie die Bänder eines Fourierspektrums sehr günstige Merkmale fuer die ASR dar.

Skizze

```
LP Koeffizienten  ->  Übertragungsfunktion  ->  Übertragungsfunktion
                      des inversen Filters          Vokaltrakt
```

Der Vorteil von aus der LP abgeleiteten spektralen Merkmalen gegenüber dem normalen Spektrum ist die Tatsache, dass sich in der Filterfunktion keine periodischen Anteile der glottalen Anregung mehr enthalten sind (denn diese wurden mit dem Fehlersignal $e(n)$ herausgefiltert. D.h. Sprecherspezifika wie Grundfrequenz, Klang der Stimme, etc. 'stören' nicht mehr.

In der Praxis hat sich eine Ordnung von mindestens $N = 12$ als günstiger Wert erwiesen, so dass man für jedes Analysefenster 12 Merkmale erhält.

Die lineare Prädiktion wurde nicht für die Spracherkennung entwickelt, sondern für die Sprachkodierung, bei der es nur auf eine möglichst getreue Rekonstruktion des Originalsignals ankommt. Das Problem der Mustererkennung mit der fehlenden Isomorphie entfällt dort.

Der Vollständigkeit halber muss hier erwähnt werden, dass es einen weiteren Weg gibt, die Übertragungsfunktion des Filters zu berechnen (siehe Abschnitt) 'Cepstrum'). Dieser andere Weg führt zu einem sehr ähnlichen Ergebnis wie die Lineare Prädiktion.

Das 'Cepstrum' (cepstrum, nur anschaulich!)

Das Cepstrum[1] eines Signalstückes wird aus dem logarithmierten Betragsspektrum dieses Signals berechnet, indem man eine Cosinus-Reihenentwicklung durchführt (*discrete cosine transformation, DCT*). Ganz analog zur Spektralberechnung, bei der das Signalstück in Sinus- und Cosinus-Anteile zerlegt wird, kann man auch das Spektrum - was ja auch nur ein digitales Signal darstellt - noch einmal transformieren.[2] Daraus erhält man die cepstralen Parameter des Signals, welche auch wieder 'Schwingungen' im Spektrum beschreiben; in diesem Zusammenhang spricht man jedoch besser von 'Welligkeit'.

[1] Der Name 'cepstrum' entsteht durch Vertauschung von 's' und 'c' in 'spectrum'.

[2] Da das Betragsspektrum immer positiv und symmetrisch zur Y-Achse liegt, reicht eine Entwicklung in Cosinus-Schwingungen.

Skizze

`Signal FFT Spektrum LOG LogSpektrum INVFFT Cepstrum (Huellkurve)`

Um das resultierende Cepstrum von einem Sprachsignal zu unterscheiden, schreibt man die Einheit der X-Achse als $1/f$ und nennt diese Dimension nicht mehr 'Zeit' sondern 'quefrency'. Im Cepstrum sind die einzelnen harmonischen Anteile eines Signals noch besser zu trennen als im Spektrum. Zum Beispiel bewirkt die Grundfrequenz und ihre harmonischen Obertöne einen starken Gipfel an der Stelle $1/f_0$ im Cepstrum.

Beispiele:

Ein Signal enthalte einen **idealen Impuls (Dirac)**. Das Spektrum dieses Impulssignals wird also vollkommen eben sein.

Skizze

`Impulssignal + Betragsspektrum`

Eine Cosinus-Reihenentwicklung dieses Spektrums wird im ersten Parameter einen Wert haben der die Höhe des Spektrum repräsentiert (die Welligkeit 0). Alle weiteren Parameter, welche die verschiedenen Welligkeiten im Spektrum beschreiben, sind logischerweise Null (weil: es gib keine Welligkeiten im Spektrum)

Skizze

Cepstrum von Dirac-Impuls

Ein Signal enthalte einen **Nasal**. Das Spektrum eines Nasals ist in erster Näherung ein breiter Gipfel bei tiefen Frequenzen.

Skizze

Skizze Nasal-Signal + Betragsspektrum

Der erste Wert des Cepstrums wird den Mittelwert der Höhe des Spektrums enthalten (die Welligkeit 0 oder auch Gleichanteil genannt). Der zweite Wert ist die Amplitude der Welligkeit, die über das ganze Spektrum eine halbe Cosinus-Schwingung ausführt. Der nächste Wert die Amplitude der Schwingung, die über das Spektrum eine komplette Cosinusschwingung ausführt, usw.

Skizze

Nasal Cepstrum

Letzterer wird den grössten Wert haben, weil das Nasalspektrum in etwa die Form einer solchen Schwingung hat (allerdings nur zur Hälfte, was den Wert wieder etwas drücken wird).

Ein Signal enthalte einen **Vokal**. Das Spektrum enthält die Grundfrequenz und deren Obertöne bei ganzzahligen Vielfachen der Grundfrequenz. Darüber ist die Formantstruktur erkennbar

Skizze

Signal + Betragsspektrum

Das Cepstrum dieses Signals wird zunächst ganz ähnlich wie im Falle des Nasals aussehen (wenn auch die Werte sicher variieren, weil die Formanten anders liegen). Bei dem Cepstralwert allerdings, der die Welligkeit der Obertöne repräsentiert, wird das Cepstrum ein deutliches Maximum haben, weil diese Schwingung im Spektrum dominant ist.

Beim Einsatz in der ASR wird das Cepstrum normalerweise nicht bis zur Schwingung der Grundfrequenz (dem Maximum) berechnet, weil man genau den Einfluss der Grundfrequenz vernachlässigen möchte. Ganz ähnlich wie die LP-Koeffizienten beschreiben also die restlichen Cepstral-Koeffizienten nur noch die 'Einhüllende', die grobe Form des Spektrums, aber nicht mehr die Linien, die durch die Glottisschwingung hervorgerufen werden.

Berechnet man das Spektrum vor der Logarithmierung nach der Mel-Skala, so spricht man von **Mel-Frequenz-Cepstral-Koeffizienten (mel frequency cepstral coefficients, MFCC)**. Diese haben eine etwas bessere Fähigkeit Sprache zu repräsentieren als das normale Ceptrum (CC) und werden daher bevorzugt in der ASR eingesetzt. Meistens werden 10-16 MFCCs berechnet und die höheren Koeffizienten (= kleine Welligkeiten im Spektrum) vernachlässigt.

```
        Praktische Übung:        anspect

        Sprachsignal sei in dog.sfs
% Es dog.sfs
  Zeitpunkt bestimmen, an dem das Spektrum berechnet werden soll: T
% anspect -c -i SP.01 -t T dog.sfs
```

Vertiefung für Interessierte

Warum berechnet man das Ceptrum aus dem logarithmierten Betragsspektrum?
Nach der Systemtheorie besteht das Sprachsignal aus dem Signal der Quelle (z.B. der Glottis) 'gefaltet' mit der Impulsantwort nachgeschalteter Filter (z.B. des Vokaltrakts, eines Telefonkanals, eines Mikrophons etc.). Die 'Faltung' ist eine komplizierte mathematische Operation, bei der eines der beiden Signale an der Y-Achse gespiegelt, anschließend über das andere Signal geschoben, die überlappenden Bereiche multipliziert und anschließend integriert werden. Wegen dieser Komplexität ist es fast unmöglich, am resultierenden Sprachsignal die beiden ursprünglichen Anteile von Quelle und Filter zu erkennen. Nach einer Transformation in den Spektralbereich wird die komplizierte Faltung allerdings zu einer einfachen Multiplikation von Anregungsspektrum und Filterübertragungsfunktion.

$$q(t) * f(t) \circ\!-\!\bullet\, Q(f)F(f)$$

Faltung im Zeitbereich entspricht Multiplikation im Spektralbereich

Logarithmiert man nun das Spektrum, so wird aus dieser Multiplikation eine Addition. Da die Fouriertransformation (oder Kosinustransformation) eine lineare Operation darstellt, bleibt diese Addition nach einer Rücktransformation des Spektrums erhalten, d.h. auch im Cepstrum sind die harmonischen Anteile Anregungssignal und Impulsantwort des Filters nicht mehr gefaltet sondern einfach addiert.

$$log[Q(f)F(f)] = log[Q(f)] + log[F(f)] \quad \bullet\!-\!\circ$$
$$\bullet\!-\!\circ \quad \mathcal{F}^{-1}\{log[Q(f)] + log[F(f)]\} = \mathcal{F}^{-1}\{log[Q(f)]\} + \mathcal{F}^{-1}\{log[F(f)]\}$$

$\mathcal{F}^{-1}\{\}$: inverse Fouriertransformation

Dadurch kann man durch einfache Subtraktion von Teilen des Cepstrums (z.B. des Gipfels der Grundfrequenz) harmonische Teile aus dem ursprünglichen Signal entfernen. Man nennt diese Operation in Anlehnung an den Begriff 'filtering' auch 'liftering'.

Warum genügt zur Rücktransformation eine Kosinustransformation?
Weil das Betragsspektrum sowohl real als auch symmetrisch zur Y-Achse ist ('gerade'). Ein solches Signal hat bei einer Fourier-Reihenzerlegung nur Kosinus-Anteile. Daher genügt es statt der vollen Fouriertransformation nur die Kosinustransformation durchzuführen; man spart dabei die Hälfte der mathematischen Operationen. (Die Fourier-Reihenzerlegung eines beliebigen Signals beinhaltet die Zerlegung sowohl in Kosinus- als auch in Sinus-Teiltöne.)

Zeitliche Ableitung von Merkmalen (delta features)

Es zeigt sich, dass der absolute Wert eines Merkmals nicht so viel Information für die ASR enthält wie die *zeitliche Änderung* des Merkmals von einem Analyse-Zeitpunkt zum nächsten. Daher berechnet man routinemäßig von jedem Merkmal auch noch die 1. zeitliche Ableitung (delta feature,

velocity, 'Geschwindigkeit') und die 2. zeitliche Ableitung (delta-delta feature, acceleration, 'Beschleunigung') und fügt diese als eigene Merkmale dem Merkmalsvektor hinzu.

Skizze

```
Merkmalsvektor  ->  Merkmalsvektor erweitert
```

Da die Merkmale in Vektoren in festen Zeitabstand (z.B. 10msec) über der Zeitachse angeordnet sind, berechnet sich die erste Ableitung in erster Näherung als einfache Differenz der vorangegangenen und nachfolgenden Merkmalswerte. Es gibt in der Literatur und Praxis mehrere sog. Regressionsformeln für die näherungsweise Berechnung der 1. Ableitung:

Sei $m(k)$ der k-te Merkmalsvektor, dann berechnet sich die erste Ableitung zu

$m'(k) = \sum_{l=-L}^{+L} l m(k+l)$ ('Butterfly-Window' der Breite $2L$)

$m'(k) = \frac{\sum_{l=1}^{L} l[m(k+l)-m(k-l)]}{2\sum_{l=1}^{L} l^2}$ (Regression)

$m'(k) = \frac{m(k+L)-m(k-L)}{2L}$ (Einfache Differenz im Abstand L)

Dabei bestimmt der Faktor L die 'Breite' des Analysefensters aus dem die Ableitung berechnet wird. In der Praxis wird L meistens zu 2 gesetzt, z.B. 'Butterfly':

$m'(k) = -2m(k-2) - m(k-1) + m(k+1) + 2m(k+2)$

(Beachte, dass in dieser Formel der ursprüngliche Wert $m(k)$ gar nicht mehr vorkommt; nur die benachbarten Werte.)

Welche der obigen Formeln in der Praxis verwendet wird, spielt für die Performanz des Spracherkenners kaum eine Rolle.

Die zweite Ableitung $m''(k)$ wird ganz einfach mit der gleichen Methode aus der ersten Ableitung berechnet:

$m''(k) = \sum_{l=-L}^{+L} l m'(k+l)$ (2. Ableitung 'Butterfly')

Die erste Ableitung wird in der englischen Literatur oft als 'delta features' oder 'velocity' und die zweite zeitliche Ableitung als 'delta delta features' oder 'acceleration' bezeichnet.

Beispiel:
Bei einer LP 12-ter Ordnung erhält man dann 12 lp + 12 delta-lp + 12 delta-delta-lp = 36 Merkmale

Warum enthalten die zeitlichen Ableitungen mehr Informationen für den Mustererkennungsprozess? Eigentlich ist dies ein Widerspruch, da sich die erste und zweite Ableitung ganz einfach aus den

primären Merkmalen berechnen lassen. Folglich müsste auch der Mustererkennungs-Algorithmus dazu in der Lage sein, wenn dies einen Vorteil erbringt.

Der Grund hierfür liegt in der Struktur von ASR-Systemen. Die meisten (nicht alle!) Ansätze verarbeiten den Input, der von der Merkmalsextraktion geliefert wird, 'frame by frame', d.h. immer ein Merkmalsvektor nach dem anderen. Daraus folgt logischerweise, dass zeitliche Beziehungen zwischen aufeinander folgenden Frames nur indirekt über die zeitliche Modellierung im Mustererkenner erfasst werden können. Wie wir noch sehen werden, verwenden jedoch die gängigen Ansätze der Mustererkennung zeitliche Modellierungen, die vom Inhalt der Merkmalsvektoren unabhängig sind. Daraus folgt, dass der Mustererkenner niemals in der Lage ist, so etwas wie eine zeitliche Ableitung in der Merkmalsvektorfolge zu berechnen. Ausnahmen von dieser Regel sind Ansätze mit neuronalen Netzen, welche häufig nicht nur einen Frame sondern auch noch deren Vorgänger-Frames und Nachfolger-Frames als Input einlesen. Konsequenterweise verzichtet man bei diesem Ansätzen auch auf zeitliche Ableitungen in den Merkmalen.

Systematische Untersuchungen, welche Merkmale am meisten zum Erkennungsprozess beitragen (durch Weglassen von einzelnen Merkmalen), haben ergeben, dass die erste Ableitung tatsächlich den größten 'Informationsanteil' zur Mustererkennung liefert. In manchen ASR-Systemen wird daher sogar auf die weitere Verarbeitung der normalen Merkmale verzichtet und man verwendet nur die erste Ableitung.

Es gibt auch eine physiologische Begründung für die Verwendung von zeitlichen Ableitungen. Aus Perzeptionsexperimenten ist bekannt, dass Übergänge zwischen Lauten ('transients') mehr Information zur Erkennung von Sprache beitragen als stationäre Teile des Sprachsignals, d.h. Teile, in welchen sich das Signal kaum ändert.