Einführung in die Signalverarbeitung

Phonetik und Sprachverarbeitung, 2. Fachsemester, Block Sprachtechnologie I

Florian Schiel

Institut für Phonetik und Sprachverarbeitung, LMU München

Signalverarbeitung - Teil 6

Allgemeines

- Unterrichtssprache ist Deutsch (englische Fachbegriffe in Klammern)
- Fragen am besten sofort; besser einmal zuviel gefragt
- Literatur:
  - Jurafsky D, Martin J H (2000): Speech and Language Processing. Prentice Hall, Kap I.7.
  - Schrüfer E (1980): Signalverarbeitung
  - Pfister B, Kaufmann T (2008): Sprachverarbeitung - Grundlagen und Methoden der Sprachsynthese und Spracherkennung. Springer-Verlag Berlin Heidelberg.
  - Rabiner, Lawrence R., Schafer R W (1978): Digital Processing of Speech Signals. Prentice-Hall, New Jersey, USA.
  - Hess W (1993): Digitale Filter. Teubner Studienbücher, B.G.Teubner, Stuttgart.
  - Harrington J, Cassidi St (1999): Techniques in Speech Acoustics. Kluwer Academic Publishers, Dordrecht/Boston/London.

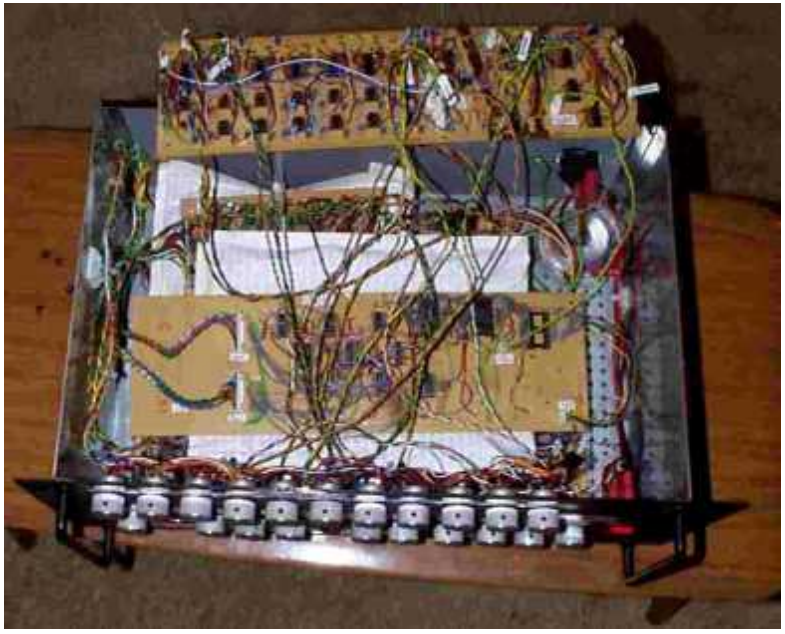
Sprachsignalverarbeitung II

Anwendung der Methoden auf die Verarbeitung von Sprache.

- Filterbänke: psychoakustisch motivierte Merkmale
- Kreuzkorrelation: Synchronisation von multiplen Signalen
- Autokorrelation: Messung von Periodizität

Digitale Filterbank

Filterbank = *traditionell: Schaltung mit parallelen Filtern*



Digitale Filterbank

Parallele Filter, die das Spektrum abdecken, realisiert entweder mit parallel laufenden digitalen Filtern oder mit DFT

Prinzip:

- führe DFT durch
- fasse gewünschte Frequenzbereiche (Bänder) durch aufaddieren der Frequenzwerte zusammen

Beispiel: Bark-Filterbank

Bark-Skala : psychoakustische Tonhöhenwahrnehmung
1 Bark = Breite einer *Frequenzgruppe* (critical band)

Zusammenhang Frequenz zu Bark:

$$z[\text{Bark}] = 13 \arctan(0.00076 \cdot f) + 3.5 \arctan((f/7500)^2)$$

Daraus ergibt sich eine Aneinanderreihung von Bändern, die jeweils 1 Bark breit sind:

0-100	100-200	200-300	300-400	400-510	510-630
630-770	770-920	920-1080	1080-1270	1270-1480	1480-1720
1720-2000	2000-2320	2320-2700	2700-3150	3170-3700	3700-4400
4400-5300	5300-6400	6400-7700	7700-9400	9400-12000	12000-15500

Beispiel: Bark-Filterbank

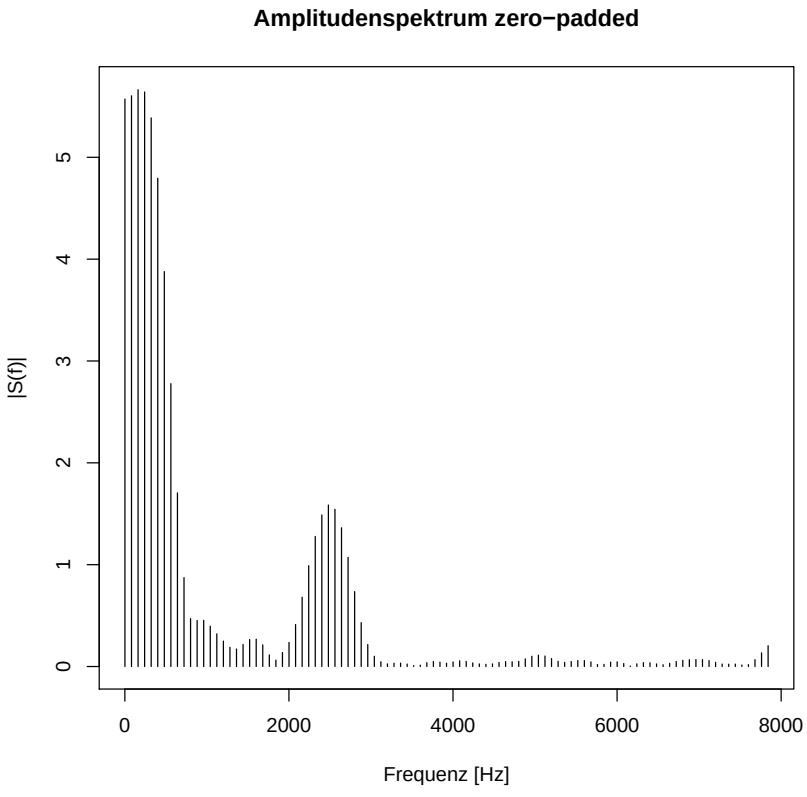
Sprachsignal $s(t_n)$ mit Abtastrate $f_{abt} = 16000Hz$
Nach Fensterung DFT mit $N = 1024$
$\rightarrow$ Abstand zweier Frequenzwerte im Spektrum $S_D(f)$:
$f_{abt}/N = 16000/1024 \approx 15Hz$

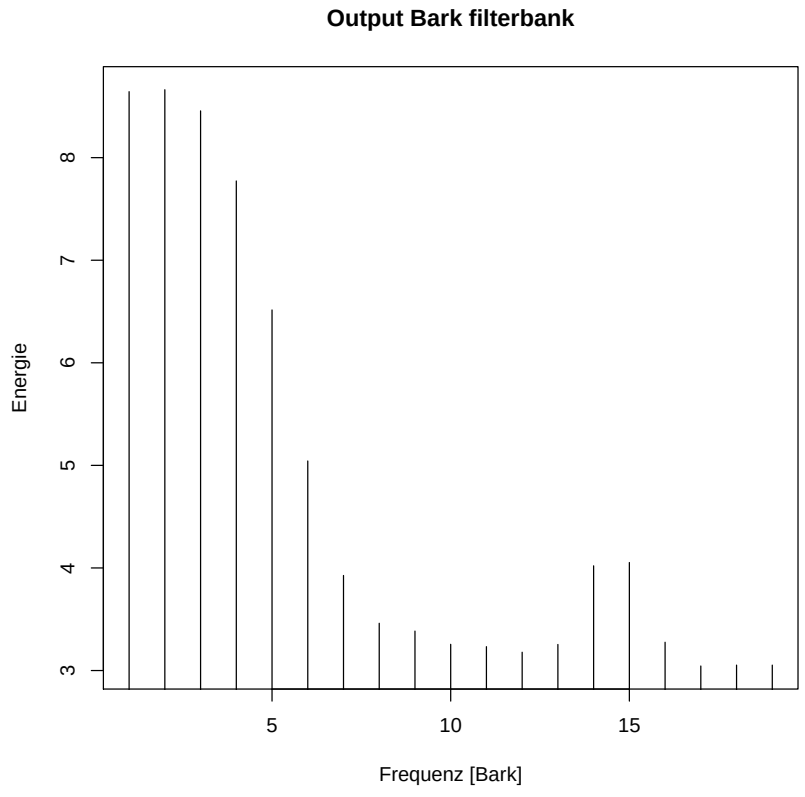
Bark-Filterbank: Alle Spektralwerte innerhalb einer Frequenzgruppe werden addiert:

$$\begin{aligned} Bark1-5 &: addiere 100/15 = 7 Linien \\ Bark6 &: addiere 120/15 = 8 Linien \\ Bark7 &: addiere 140/15 = 9 Linien \\ Bark8 &: addiere 150/15 = 10 Linien \\ Bark9 &: addiere 160/15 = 11 Linien \\ Bark10 &: addiere 190/15 = 13 Linien \\ Bark11 &: addiere 210/15 = 14 Linien \\ Bark12 &: addiere 240/15 = 16 Linien \end{aligned}$$

Beispiel: Bark-Filterbank

Vergleich normale DFT und Output Bark-Filterbank





Andere Filterbänke

Nach dem selben Prinzip

- Mel-Filterbank (z.B. je 50 mel breite Bänder)
- Halbton- oder Vollton-Filterbank
- Hochton-/Mittelton-/Tiefton-Filterbank (bestimmte Frequenzbereiche)
- ...

Kreuzkorrelation

Kreuzkorrelation zweier Signale:
Verschieben der beiden Funktion um einen Abtastwert, multiplizieren und aufaddieren $\rightarrow$ ein Wert der Kreuzkorrelation (cross correlation)

$$c(t_k) = \frac{1}{N} \sum_{n=1}^{N} s_1(t_n) s_2(t_{n+k})$$

Beachte: I.G. zur *Faltung* wird hier keine Funktion gespiegelt!

Kreuzkorrelation dient in der Phonetik oft zur *Synchronisation* von ähnlichen Signalen (z.B. von zwei Mikrophonen):
$c(t_k)$ wird nur für einen Bereich $k = -K \cdots + K$ berechnet, in dem man die Assynchronie vermutet. Dann hat $c(t_k)$ bei der tatsächlichen Assynchronie ein Maximum.

Kreuzkorrelation

Beispiel: Experiment mit Headset-Mikrophon und Video.
Headset- und Video-Ton-Signal haben gleiche Abtastrate.
Videodaten sollen mit Headset-Kanal synchronisiert werden.

$\rightarrow$ Kreuzkorrelation von Headset-Signal $h(t_n)$ mit Tonsignal der Kamera $v(t_n)$ der ersten 10 Sekunden Aufzeichnung:

$$c(t_k) = \frac{1}{N} \sum_{n=1}^{N} h(t_n) v(t_{n+k})$$

$c(t_k)$ hat Maximum bei $k = 3425$
$\rightarrow$ Video-Signal ist um 3425 Samples früher als Headset.

Einfügen von 3425 Null-Samples am Anfang des Headset-Signals $\rightarrow$ beide Signale synchron.

Autokorrelation

Autokorrelation = Kreuzkorrelation mit sich selbst

$$a(t_k) = \frac{1}{N} \sum_{n=1}^{N} s(t_n) s(t_{n+k})$$

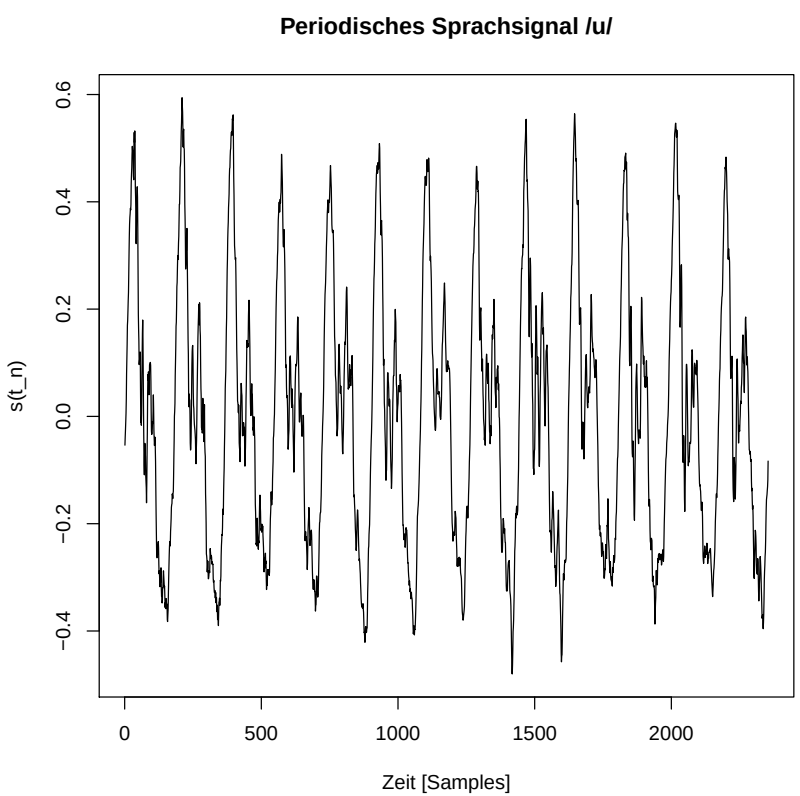
Die Autokorrelation $a(t_k)$ eines Signals $s(t_n)$ dient oft dazu, Periodizitäten in $s(t_n)$ zu finden:
$a(t_k)$ hat bei dem t_k ein erstes Maximum, das der Länge der Grundschwingung (T_0) einer periodischen Schwingung entspricht.

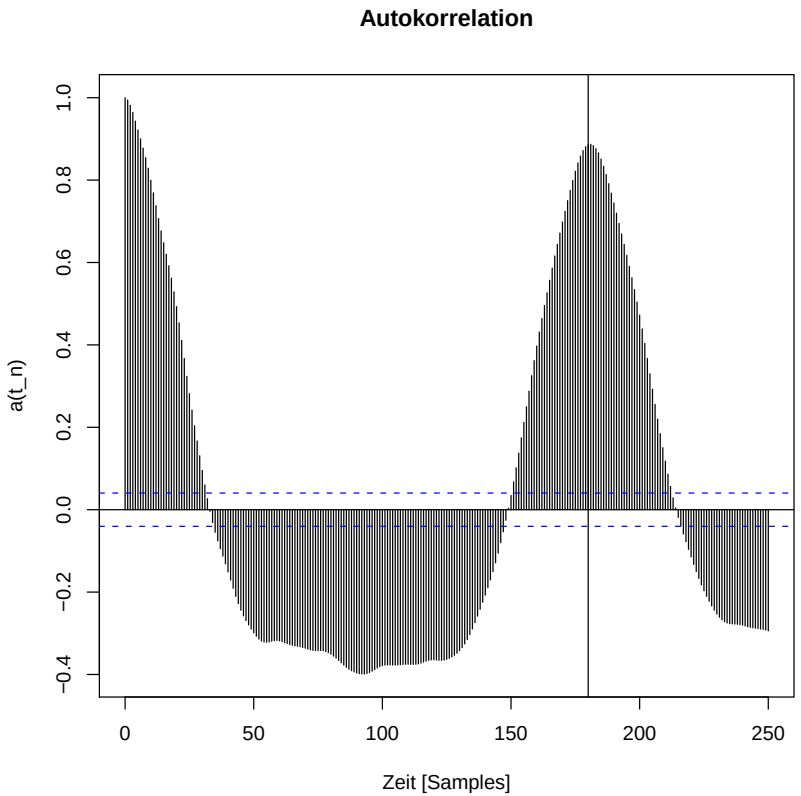
$\rightarrow$ Grundfrequenz-Detektor

Die Fouriertransformierte von $a(t_k)$ ist das Leistungsspektrum von $s(t_n)$: $\mathcal{F}\{a(t_k)\} = S_D^2(f)$

Autokorrelation

Beispiel: Grundfrequenz eines /u/-Lauts bestimmen
Abtastrate des Signals $f_{abt} = 22050 Hz$





Autokorrelation

Das autokorrelierte Signal hat ein erstes Maximum bei ca. 180 Samples (senkrechter Strich).

Eine Messung der Periodizität im Sprachsignal ergab einen zeitlichen Abstand zwischen zwei Glottisschlägen von $T_0 = 0.008$

Kontrolle Periodendauer in Samples N_0:

$$N_0 = T_0 \cdot f_{abt} = 0.008 sec \cdot 22050 Hz = 176 \approx 180 \text{Samples}$$

Fragen

Es gibt zwei Möglichkeiten eine Filterbank zu realisieren. Welche?

Wie groß ist der Abstand zweier Frequenzwerte einer DFT mit 1000 Werten angewandt auf ein Signal mit Abtastrate 20000Hz?

Zur nachträglichen Synchronisation zweier Signale verwendet man die Autokorrelation oder die Kreuzkorrelation?

Ist die Autokorrelierte eines stimmlosen Frikativs
- ein periodisches Signal?
- ein stochastisches Signal?
- ein Signal mit periodischen Maxima?

Wie könnte man mit Hilfe der Autokorrelation stimmhafte von stimmlosen Frikativen unterscheiden?