

Spracherkennung

Thema 3: Merkmalsberechnung

WS 2017/18

Peter Birkholz

Institut für Akustik und Sprachkommunikation, TU Dresden

Motivation

Als Merkmale für die automatische Spracherkennung sind die **Mel Frequency Cepstral Coefficients (MFCC)** als de-facto Standard etabliert. Sie lassen sich effizient berechnen und berücksichtigen elementare Eigenschaften unseres Gehörs:

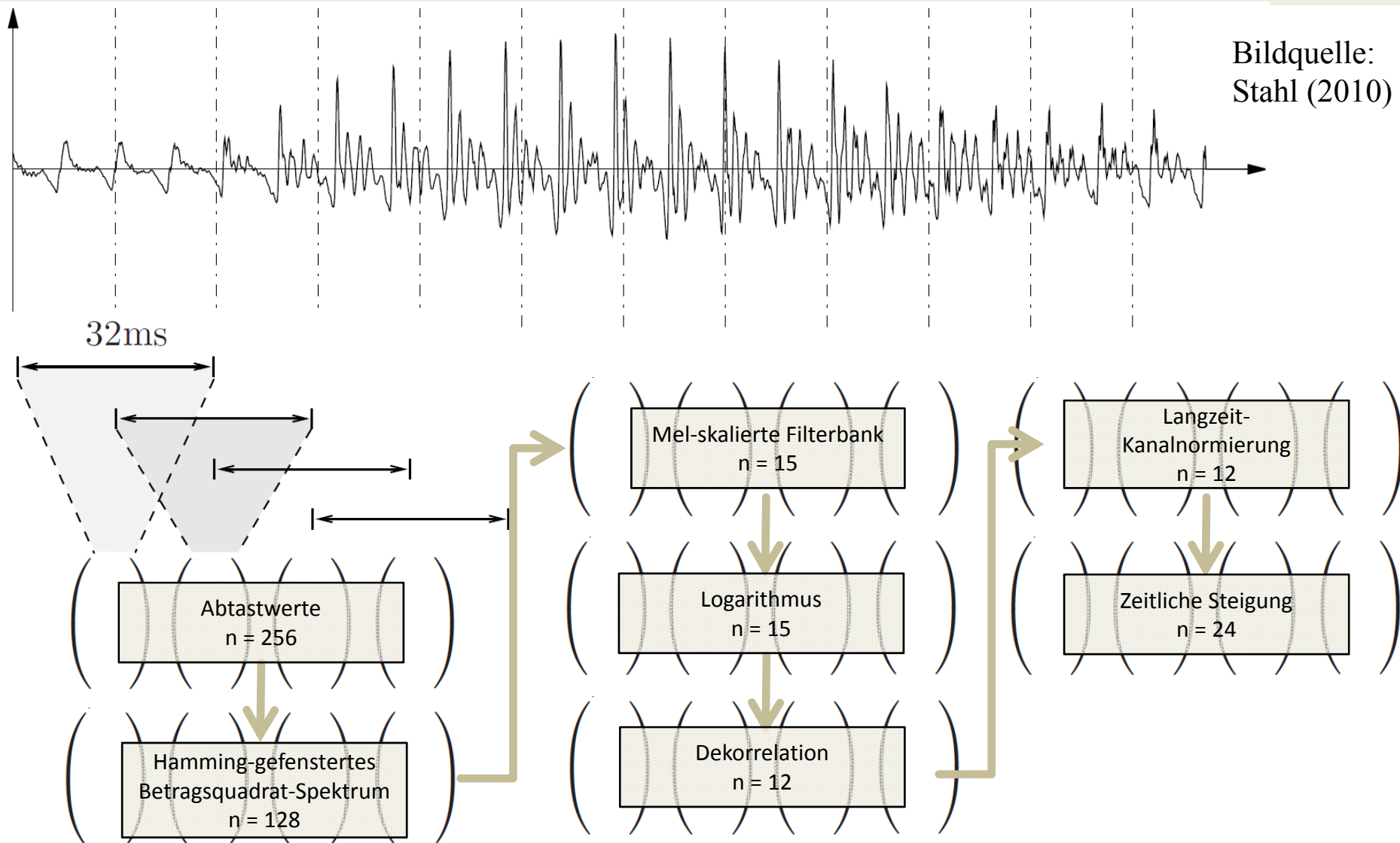
- logarithmische Wahrnehmung der Amplitude
- geringere Frequenzselektivität bei hohen Frequenzen

Die MFCC werden jeweils über einem kurzen Signalabschnitt (**Frame**) berechnet (Länge z. B. 32 ms) und bilden einen **Merkmalsvektor** für jeden Frame.

Die Veränderlichkeit des Sprachsignals wird durch eine Folge überlappender Frames erfasst. Die Überlappung ergibt sich aus dem **Vorschub** (z. B. 16 ms).

Schritte bei der Merkmalvektorberechnung

Bildquelle:
Stahl (2010)



Warum Frequenzbereichsmerkmale?

- Ziel der Signalanalyse ist die Berechnung von Merkmalen, die sich gut für die Klassifikation eignen (gute Unterscheidung der Merkmalvektoren verschiedener Klassen).
- Eine Klassifikation auf Basis der Abtastwerte des Zeitsignals würde schlecht funktionieren. Im Zeitbereich sind z. B. zwei (gleich klingende) Sinustöne mit der gleichen Frequenz und Amplitude, aber 180° Phasendifferenz, sehr unterschiedlich.
- Wdh.: Die Phaseninformation der Teiltöne spielt für (quasi-) periodische Signale kaum eine Rolle
- Unser Innenohr macht eine Frequenzzzerlegung der einfallenden Schalle.
- Die feine Zeitstruktur ist gerade bei Plosivlauten auch wichtig, wird mit den MFCC aber ignoriert.

Framelänge

Was ist die beste Länge für einen Frame?

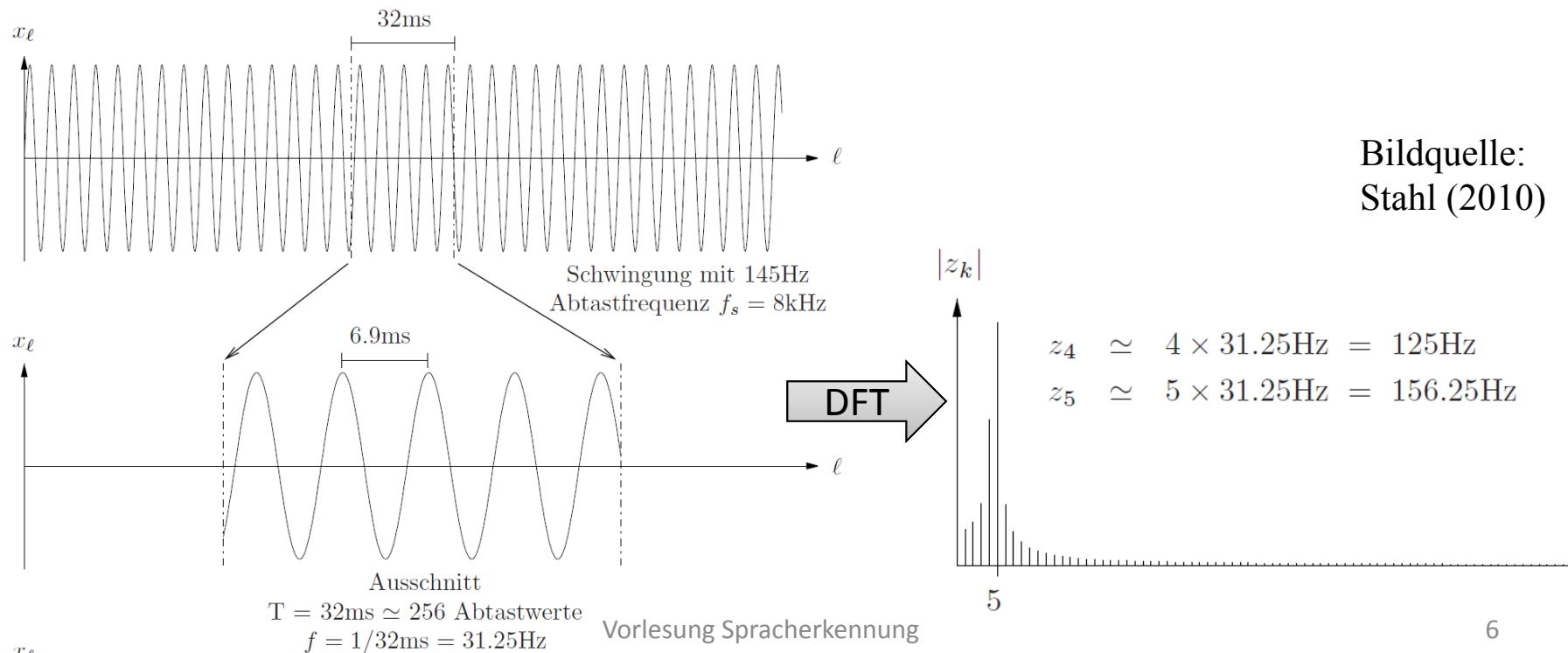
- Ein Frame sollte möglichst kurz sein, damit schnelle Änderungen im Signal gut erfasst werden. Idealerweise möchte man die Frequenzzusammensetzung an einem *Zeitpunkt* berechnen.
- Ein Frame sollte hinreichend lang sein, damit das Spektrum gut aufgelöst wird und invariant gegenüber der relativen Lage bezüglich der Grundperioden ist.
- Demonstration der Einflüsse auf das Kurzzeitspektrum von
 - Framelänge (Breitband- vs. Schmalbandspektrum)
 - Frameposition im Verhältnis zu den Perioden (Glottisverschlusszeitpunkten)

Fensterung

Warum wird das Zeitsignal vor der DFT gefenstert?

Das Signal im Frame ist i.d.R. nicht genau T -periodisch (mit der Framedauer T). Die DFT geht jedoch von einer periodischen Fortsetzung des Signals aus.

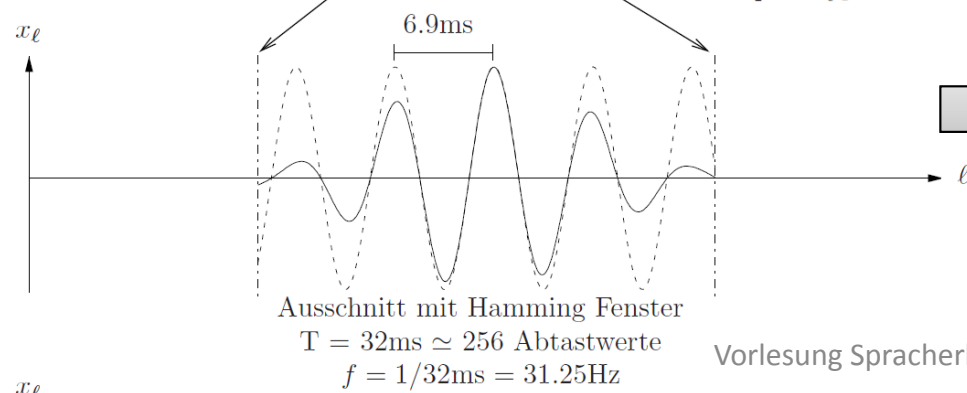
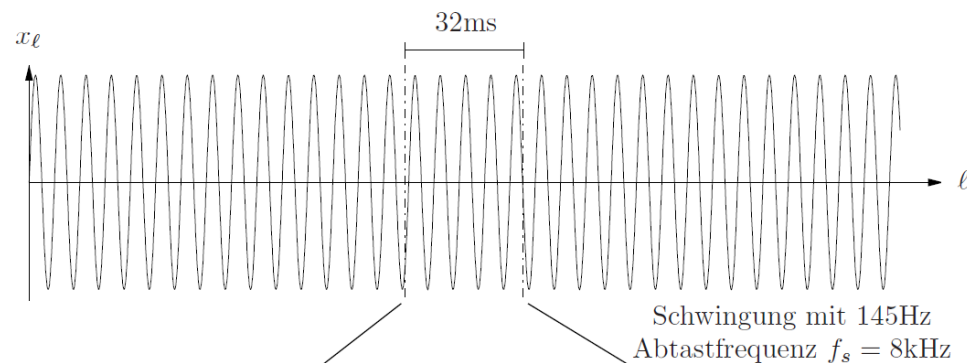
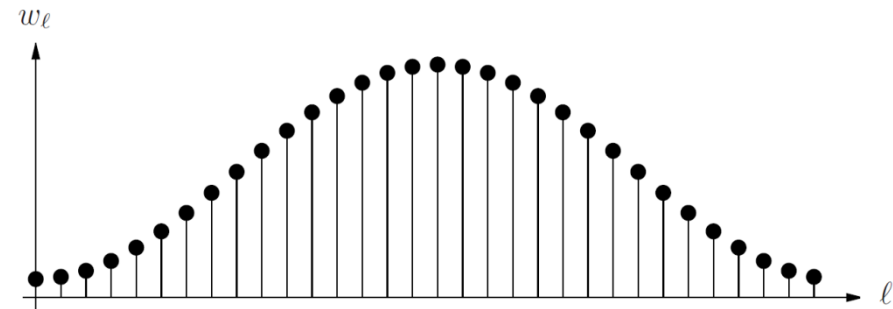
Frame aus einem Sinuston ohne Fensterung:



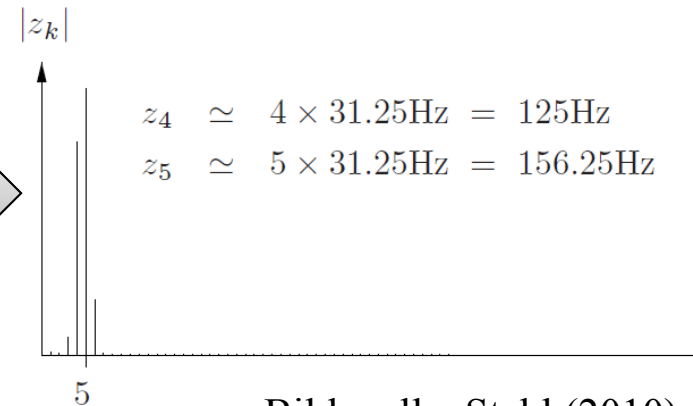
Fensterung

Sehr gebräuchlich ist das
Hamming-Fenster:

$$w_\ell = 0.54 - 0.46 \cos(2\pi\ell/n)$$



DFT



Bildquelle: Stahl (2010)

Betragsquadrat-Spektrum (Leistungsdichte-Spektrum)

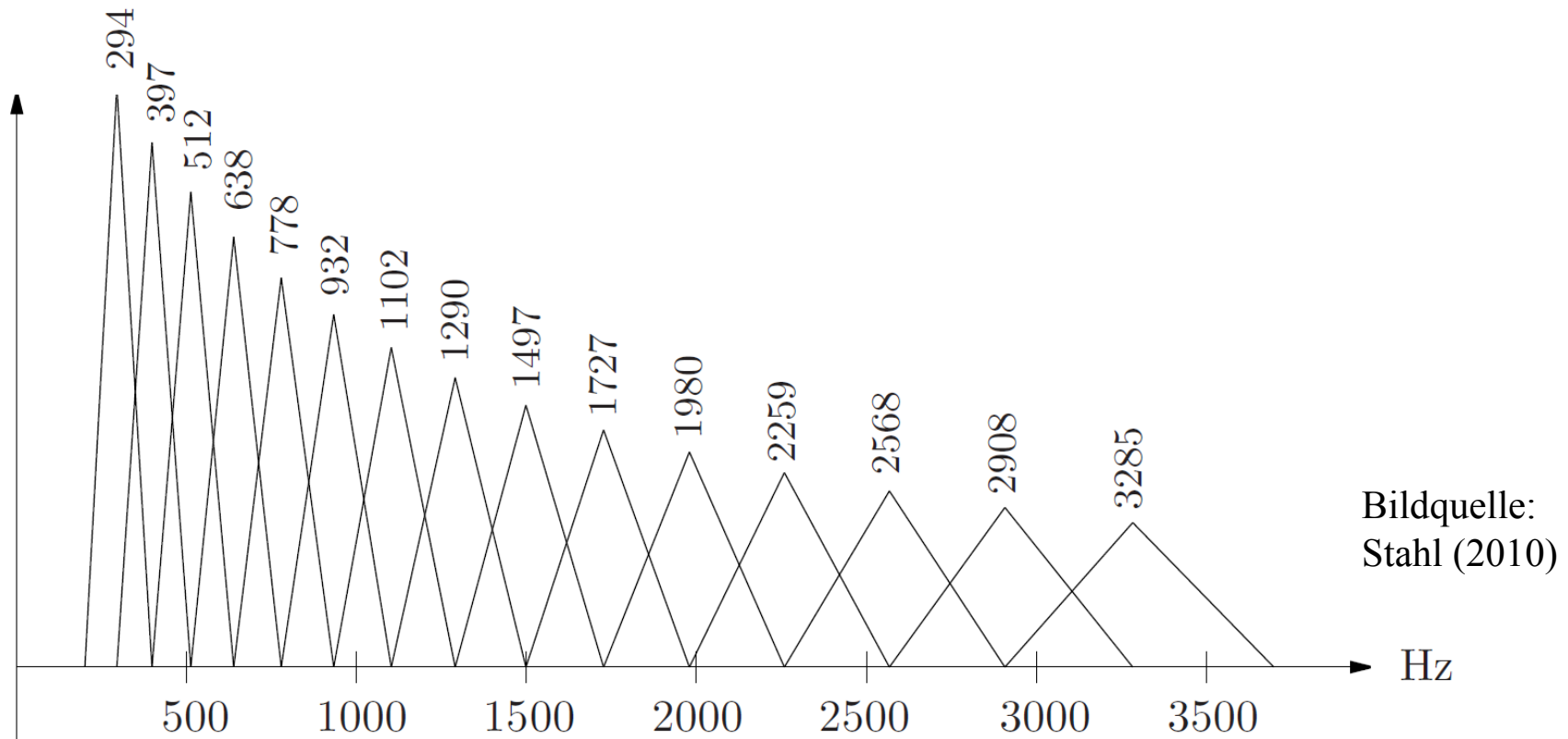
- Unser Frame sei 256 Abtastwerte lang (=32 ms Framelänge bei einer Abtastrate von 8 kHz). Die gefensterte DFT dieses Frames ergibt 256 komplexwertige Fourier-Koeffizienten z_k . Da die Koeffizienten in konjugiert komplexen Paaren auftauchen, brauchen wir nur die ersten 128 davon behalten.
- Von jedem Koeffizienten wird das Betragsquadrat gebildet:
$$z'_k = |z_k|^2 = \text{Re}\{z_k\}^2 + \text{Im}\{z_k\}^2$$
- Man könnte alternativ auch den Betrag bilden, spart sich aber mit dem *Betragsquadrat* eine Wurzeloperation.
- Durch die Betragsbildung geht die (irrelevante) Phaseninformation verloren (=Informationsreduktion).

Mel-skalierte Filterbank

- Die Auflösung des Frequenzbereichs von 0-4 kHz in 128 Frequenzkomponenten ist für die Spracherkennung nicht erforderlich. Es reichen bereits 15 Koeffizienten sehr gut aus.
- Daher: Unterteilung des Bereichs von 0-4 kHz in 15 Intervalle und Berechnung der mittleren Amplitude in jedem Intervall → 15-dimensionale Merkmalvektoren.
- Da das Ohr tiefe Frequenzen besser auflösen kann als hohe, macht man die 15 Intervalle unterschiedlich breit (zunehmend breitere Intervalle zu höheren Frequenzen).
- Auch macht man die Intervalle überlappend und gewichtet die Frequenzen innerhalb eines Intervalls mit einer Dreiecksfunktion.

Mel-skalierte Filterbank

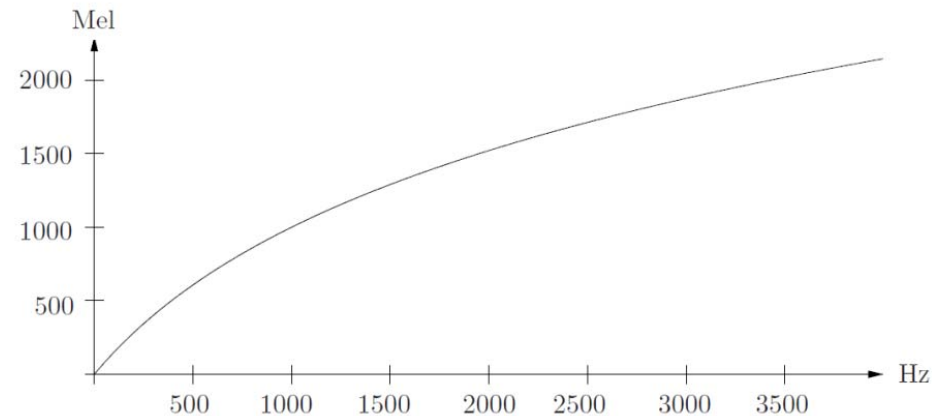
- Die Frequenzbänder werden so gewählt, dass sie auf der *Mel-Skala* (psychoakustische Skala) gleich breit sind.
- Der Flächeninhalt unter jedem Dreieck soll Eins sein.



Bildquelle:
Stahl (2010)

Mel-skalierte Filterbank

$$f_{\text{Hz}} = 700(10^{f_{\text{Mel}}/2595} - 1)$$
$$f_{\text{Mel}} = 2595 \log_{10}(f_{\text{Hz}}/700 + 1)$$



Wir treffen folgende Entscheidungen:

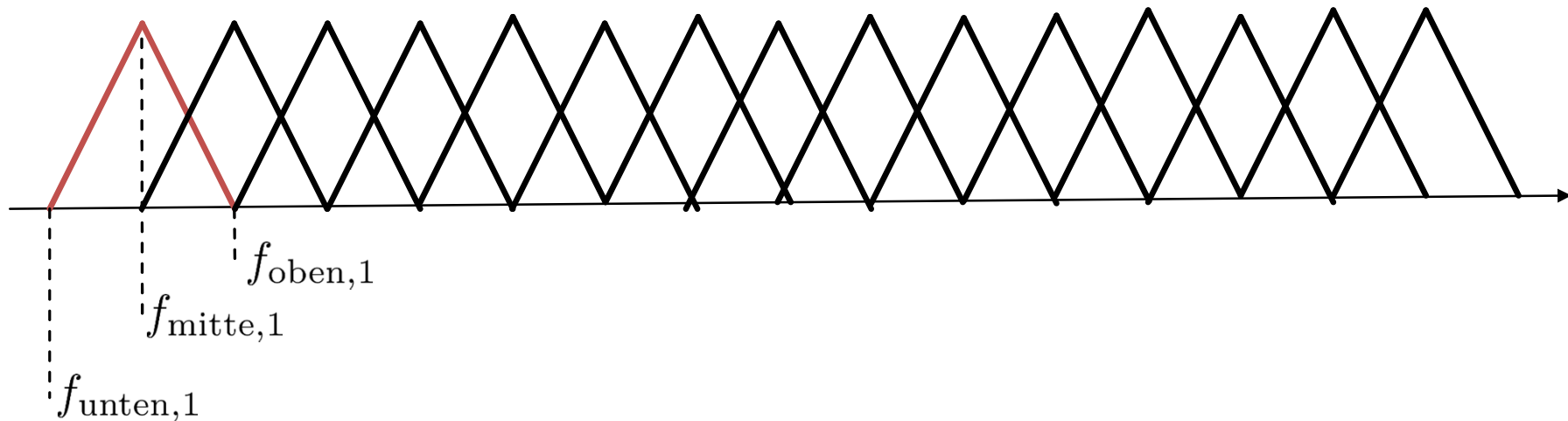
- Beginn des untersten Frequenzbandes bei 200 Hz (darunter sind nur Rauschen, Netzbrummen, Sprachgrundfrequenz)
- Oberstes Frequenzband endet bei 3700 Hz (Bereich darüber ist bei 8 kHz Abtastrate stark vom Antialiasing-Filter der Soundkarte beeinflusst)

Der Bereich 200-3700 Hz entspricht 283-2072 Mel.

Mel-skalierte Filterbank

17 äquidistante Frequenzen auf der Mel-Skala (ca. 112 Mel Abstand):

283	295	507	619	731	843	955	1067	1179	1291	1403	1515	1627	1739	1851	1962	2072	Mel
200	294	397	512	3285	3700	Hz



Höhe jedes Dreiecks so wählen, dass die Fläche Eins wird:

$$(f_{\text{oben},i} - f_{\text{unten},i}) \cdot h_i / 2 = 1 \quad \rightarrow \quad h_1 = 0.01015; h_2 = 0.00917; \dots$$

Mel-skalierte Filterbank

Gewichtung der Fourierkoeffizienten mit der Frequenz f in der linken und rechten Dreieckshälfte mit (Index i weggelassen):

$$g_{\text{links}} = (f - f_{\text{unten}}) \cdot \frac{h}{f_{\text{mitte}} - f_{\text{unten}}} \quad \text{und} \quad g_{\text{rechts}} = -(f - f_{\text{mitte}}) \cdot \frac{h}{f_{\text{oben}} - f_{\text{mitte}}}$$

Im Bereich des ersten Frequenzbandes von 200-397 Hz liegen z.B. die Fourier-Koeffizienten 7 bis 12 mit den Frequenzen

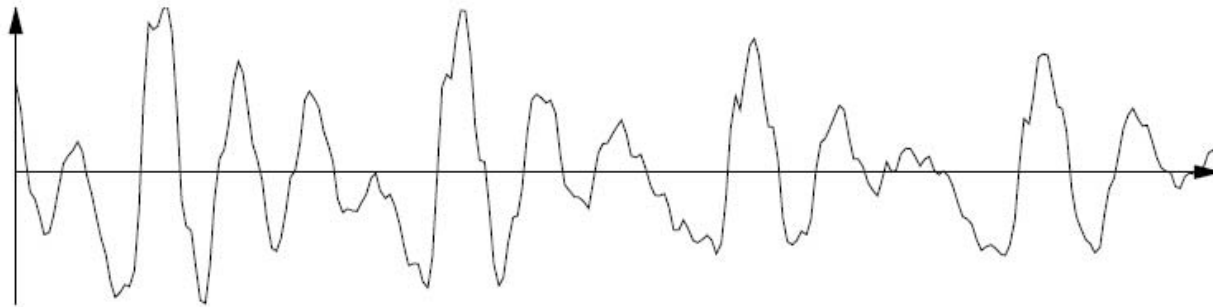
$$7 \times 31.25\text{Hz}, \dots, 12 \times 31.25\text{Hz}.$$

Diese werden mit den Gewichten

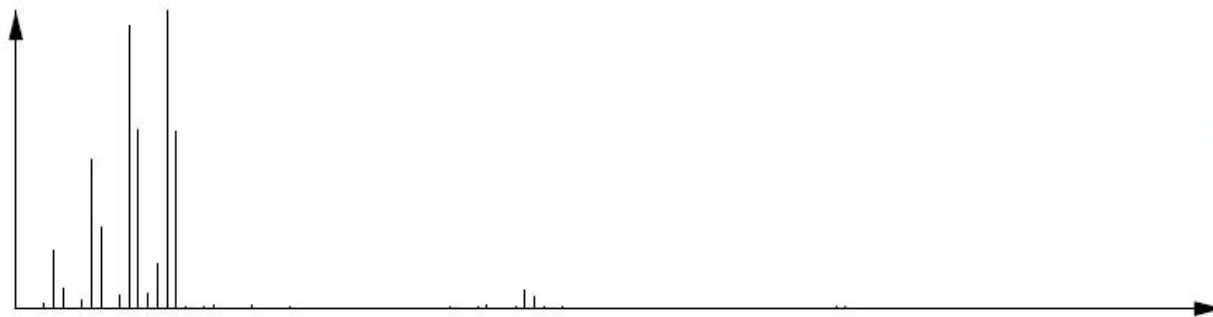
$$0.00202, 0.00540, 0.00877, 0.00830, 0.00525, 0.00220$$

multipliziert und aufsummiert, um die mittlere Leistung im ersten Band zu berechnen. Analog wird mit den restlichen Bändern verfahren. Durch die Mittelung *wird effektiv die spektrale Feinstruktur (besonders durch F0) entfernt*.

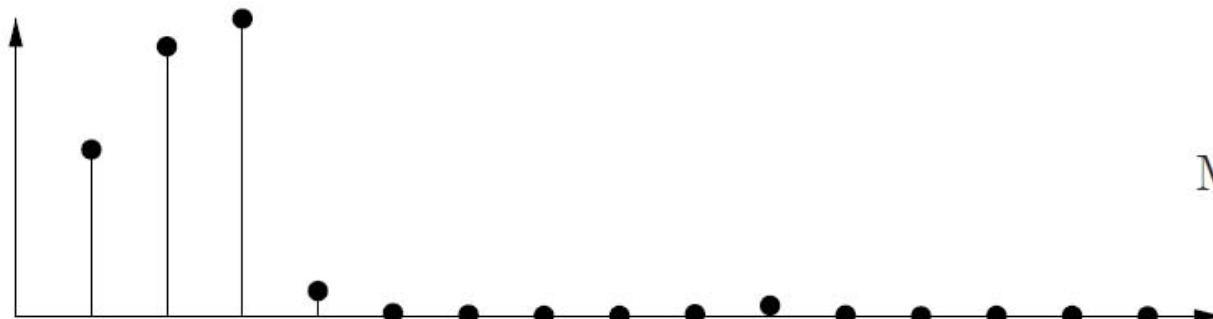
Mel-skalierte Filterbank



Abtastwerte
 $n = 256$



Hamming gefenstertes
Betragsquadrat Spektrum
 $n = 128$



Mel skalierte Filterbank
 $n = 15$

Logarithmierung

Nächster Schritt: Logarithmierung aller 15 Komponenten der Filterbank-Ausgangsvektoren.

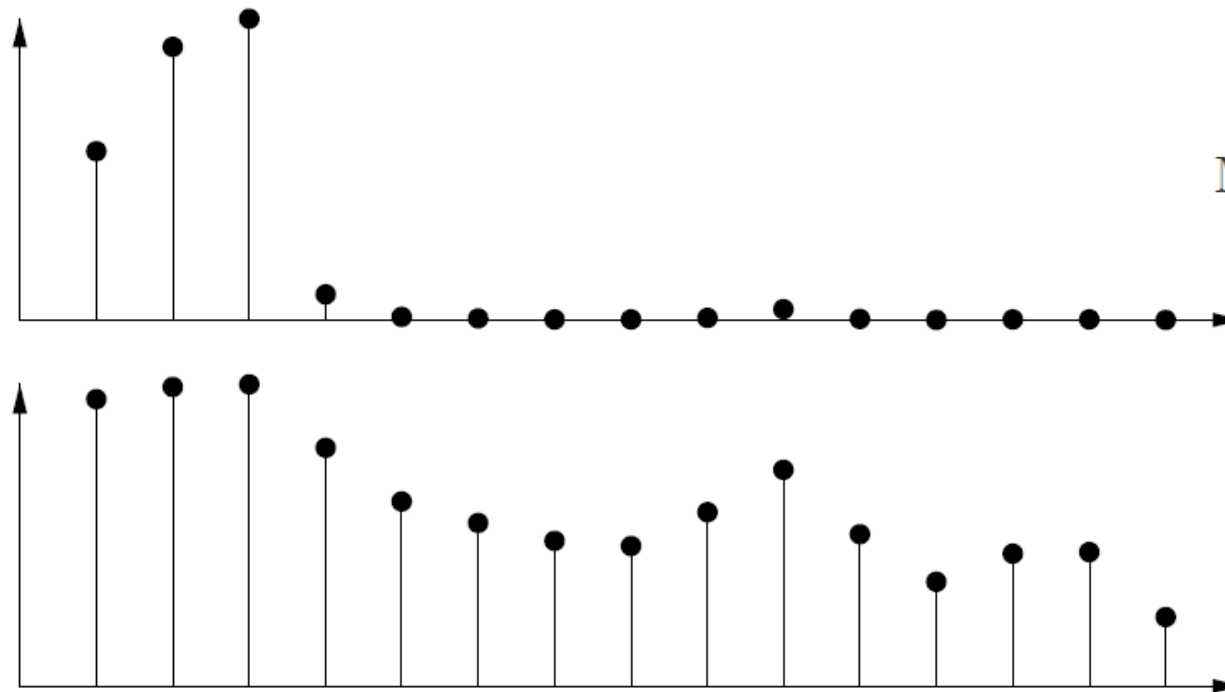
Dadurch:

- Reduzierung der enormen Größenunterschiede zwischen tiefen und hohen Frequenzen
- Lautstärkeunterschiede in der Aufnahme werden von einem konstanten Faktor zu einem konstanten Summanden. Der Summand wird bei der Langzeit-Kanalnormierung „entfernt“.

$$\log(ax_i) = \log(a) + \log(x_i), \quad i = 1, \dots, 15$$

Bei (fast) Stille kann die Logarithmierung große negative Werte ergeben. Daher wird oft $\log(x + c)$ statt $\log(x)$ berechnet, wobei c eine kleine positive Konstante ist.

Logarithmierung



Mel skalierte Filterbank
 $n = 15$

Logarithmus
 $n = 15$

Bildquelle: Stahl (2010)

Dekorrelation durch diskrete Kosinustransformation (DCT)

- Benachbarte Vektorkomponenten im mel-skalierten und logarithmierten Spektrum haben meist ähnliche Werte; sie sind statistisch voneinander abhängig (korreliert).
- Ein Basiswechsel kann die Grobinformation und Detailinformation im Spektrum trennen → Komprimierung der Information von 15 auf 12 Vektorkomponenten!

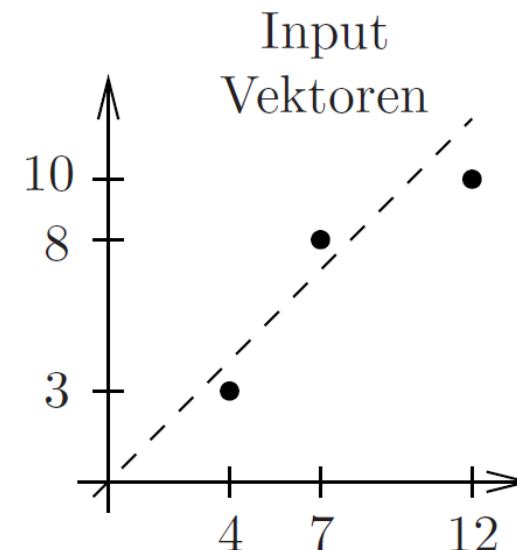
Beispiel für den 2D-Fall: Gegeben seien die Eingangsvektoren

$$\begin{pmatrix} 4 \\ 3 \end{pmatrix}, \begin{pmatrix} 7 \\ 8 \end{pmatrix}, \begin{pmatrix} 12 \\ 10 \end{pmatrix}$$

die wir nun bezüglich der neuen Basis

$$B = \left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right)$$

darstellen.



Dekorrelation durch diskrete Kosinustransformation (DCT)

Die Darstellung bezüglich der neuen Basis ist:

$$\begin{aligned}\begin{pmatrix} 4 \\ 3 \end{pmatrix} &= 3.5 \begin{pmatrix} 1 \\ 1 \end{pmatrix} + 0.5 \begin{pmatrix} 1 \\ -1 \end{pmatrix} \\ \begin{pmatrix} 7 \\ 8 \end{pmatrix} &= 7.5 \begin{pmatrix} 1 \\ 1 \end{pmatrix} - 0.5 \begin{pmatrix} 1 \\ -1 \end{pmatrix} \Rightarrow \begin{pmatrix} 3.5 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 7.5 \\ -0.5 \end{pmatrix}, \begin{pmatrix} 11 \\ 1 \end{pmatrix} \\ \begin{pmatrix} 12 \\ 10 \end{pmatrix} &= 11 \begin{pmatrix} 1 \\ 1 \end{pmatrix} + 1 \begin{pmatrix} 1 \\ -1 \end{pmatrix}\end{aligned}$$

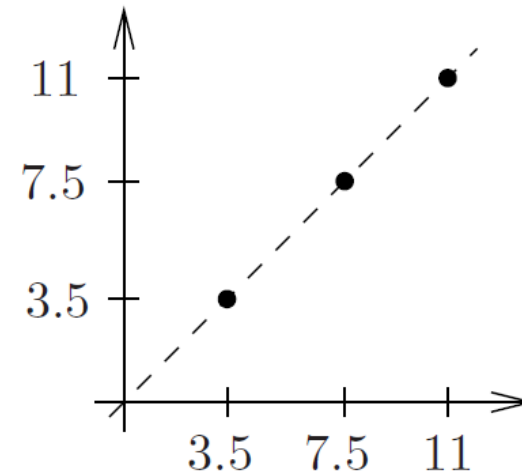
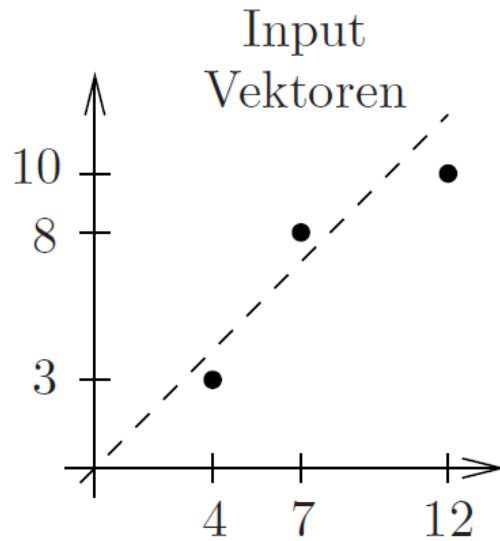
Die wichtige Information steckt in der ersten Komponente, während die zweite Komponente nur kleine Werte enthält, die wir vernachlässigen. Wir erhalten die 1D-Ergebnisvektoren

$$(3.5), (7.5), (11)$$

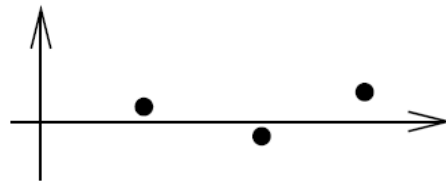
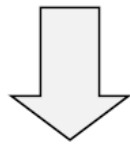
Rekonstruktion der Originalvektoren gelingt in guter Näherung ...

Dekorrelation durch diskrete Kosinustransformation (DCT)

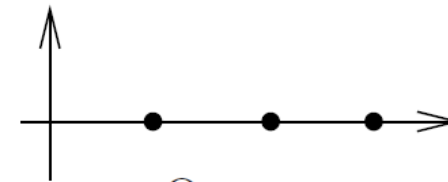
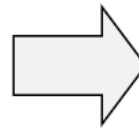
Korrelation bei
Merkmalen kann
u.a. für Machine
Learning
Algorithmen
problematisch
sein -> daher
dekorrelieren



Basiswechsel
(Dekorrelation)



zweite Komponente
vernachlässigen



Output
Vektoren



Rekonstruktion

Dekorrelation durch diskrete Kosinustransformation (DCT)

Die *DCT* verwendet als neue n -dimensionale Basis die Vektoren:

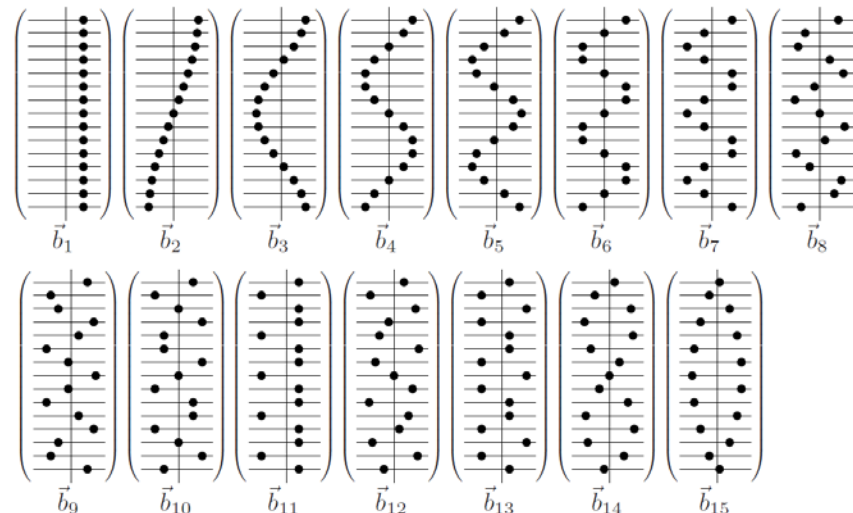
$$\vec{b}_1 = \sqrt{1/n} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$\vec{b}_i = \sqrt{2/n} \begin{pmatrix} \cos\left(\left(1 - \frac{1}{2}\right) \pi(i-1)/n\right) \\ \cos\left(\left(2 - \frac{1}{2}\right) \pi(i-1)/n\right) \\ \vdots \\ \cos\left(\left(n - \frac{1}{2}\right) \pi(i-1)/n\right) \end{pmatrix}, i = 2, 3, \dots, n$$

Neue Basis:

$$B = (\vec{b}_1, \vec{b}_2, \dots, \vec{b}_n)$$

Basisvektoren für $n=15$:



Bildquelle:
Stahl (2010)

Dekorrelation durch diskrete Kosinustransformation (DCT)

Da es sich bei B um eine Orthonormalbasis handelt, kann ein Vektor \vec{x} bzgl. der alten Basis wie folgt in einen Vektor \vec{y} bzgl. der neuen Basis B transformiert werden:

$$\vec{y} = B^T \vec{x} \quad (\text{allgemein: } \vec{y} = B^{-1} \vec{x})$$

Daraus folgt:

$$y_1 = \sqrt{1/n} \sum_{j=1}^n x_j$$

$$y_i = \sqrt{2/n} \sum_{j=1}^n x_j \cos \left(\left(j - \frac{1}{2} \right) \pi (i-1)/n \right), \quad i = 2, 3, \dots, n$$

Die letzten drei Komponenten der resultierenden Vektoren werden verworfen. Die erste Komponente enthält ein Maß für die Energie.

- Die 15-dim. Vektoren wurden auf 12-dim. Vektoren reduziert.
- Die Vektorkomponenten sind nun unkorreliert (positiv für statistische Modellierung)

Langzeit-Kanalnormierung

Die Signale, die wir bei der Erkennung miteinander vergleichen wollen, wurden i. d. R. nicht mit dem gleichen Mikrofon, der gleichen Verstärkung usw. aufgenommen. Dies führt u.a. zu *Unterschieden in der Signalverstärkung*, die oft auch noch *frequenzabhängig* sind. Die Beseitigung der variablen Verstärkung ist das Ziel der Langzeit-Kanalnormierung.

Wenn die konkrete Verstärkung des „reinen“ Signals in den Frequenzbändern der Mel-Filterbank mit dem Vektor \vec{a} bezeichnet wird, und der Ausgang der Filterbank zum Zeitpunkt t *ohne* Verstärkung als $\vec{s}^{(t)}$, dann ist das vorliegende Signal $\vec{x}^{(t)}$ am Filterausgang $\vec{x}^{(t)} = \vec{a}^2 \cdot \vec{s}^{(t)}$.

(Quadrierung wg. des Leistungsdichte-Spektrums)

Langzeit-Kanalnormierung


nach Logarithmierung:

$$\log(\vec{x}^{(t)}) = 2 \cdot \log(\vec{a}) + \log(\vec{s}^{(t)})$$

nach DCT:

$$\text{DCT}[\log(\vec{x}^{(t)})] = 2 \cdot \text{DCT}[\log(\vec{a})] + \text{DCT}[\log(\vec{s}^{(t)})]$$

bzw.


$$\vec{y}^{(t)} = \vec{a}' + \vec{s}'^{(t)}$$

Mittelwertbildung über viele Frames (z. B. 1 Sekunde):

$$\underline{\vec{y}} = \underline{\vec{a}'} + \underline{\vec{s}'}$$

Langzeit-Kanalnormierung

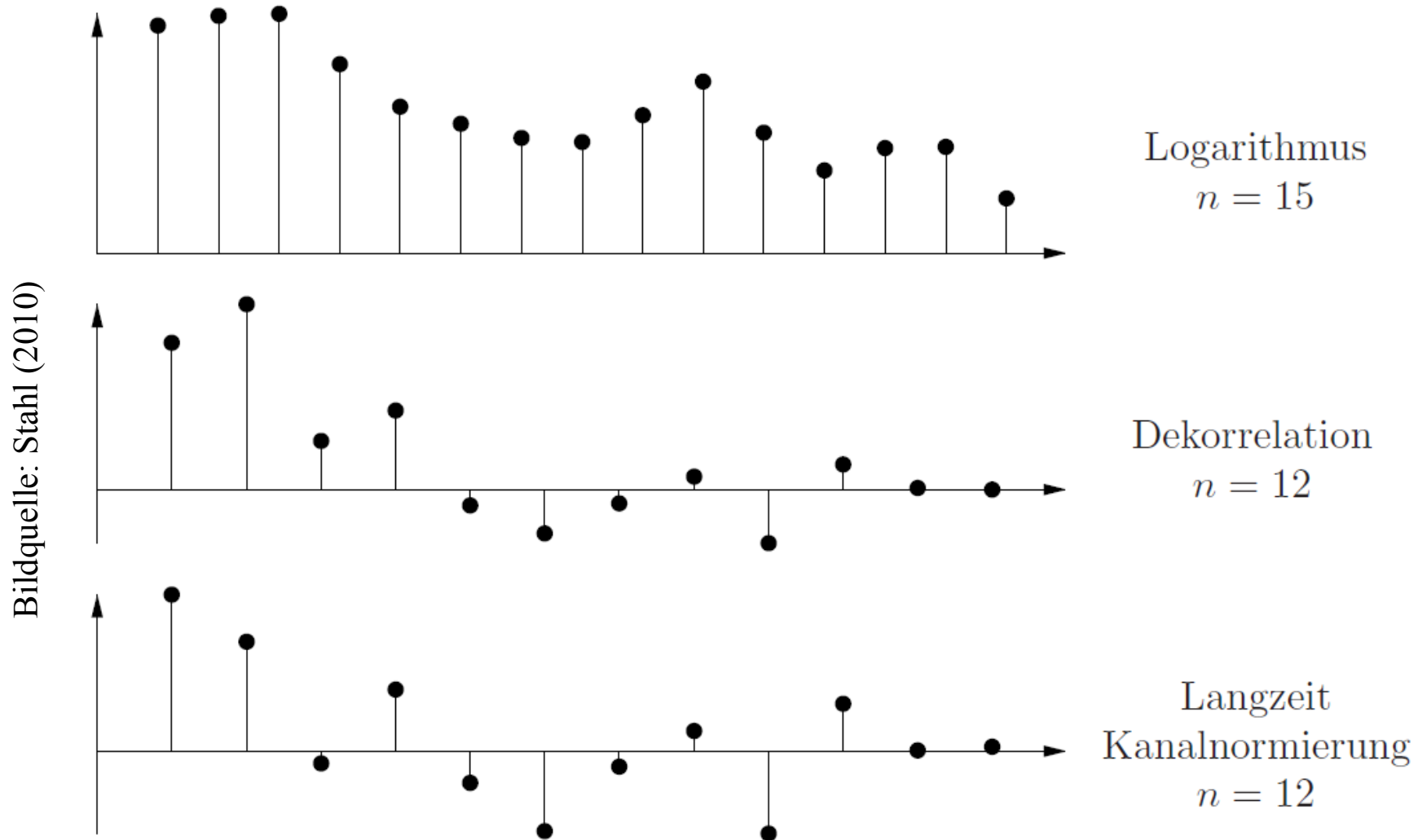
Subtraktion des Langzeit-Mittelwerts von dem aktuellen Merkmalvektor ergibt:

$$\vec{y}^{(t)} - \underline{\vec{y}} = \vec{s}^{(t)} - \underline{\vec{s}} \rightarrow \text{die Verstärkung wurde eliminiert!}$$

Das Ergebnis der Normierung hängt lediglich von den „reinen“ Merkmalvektoren ab!

Diese Art der Normierung wird auch als „channel equalization“, „long term mean subtraction“ oder „cepstral mean subtraction“ bezeichnet.

Transformationen am Beispielvektor



Zeitliche Steigung

Der bisherige Merkmalvektor enthält keine Information über die zeitliche Veränderung des Signals. Wird es lauter oder leiser? Verschieben sich spektrale Anteile?

Daher Approximation der zeitlichen Ableitung jeder Vektorkomponente $x_i^{(t)}$ aus den entsprechenden Werten in den zwei vorangegangenen und nachfolgenden Frames:

$$\frac{1}{6} \left(-2x_i^{(t-2)} - x_i^{(t-1)} + x_i^{(t+1)} + 2x_i^{(t+2)} \right)$$

In den ersten und letzten beiden Merkmalvektoren der Folge mit der Länge m werden die fehlenden Werte für

$$x_i^{(-1)}, x_i^{(-2)}, x_i^{(m+1)} \text{ und } x_i^{(m+2)}, \quad i = 1, \dots, 12$$

durch Null ersetzt.

Hinzufügen der 12 Ableitungen zum bisherigen Merkmalvektor.

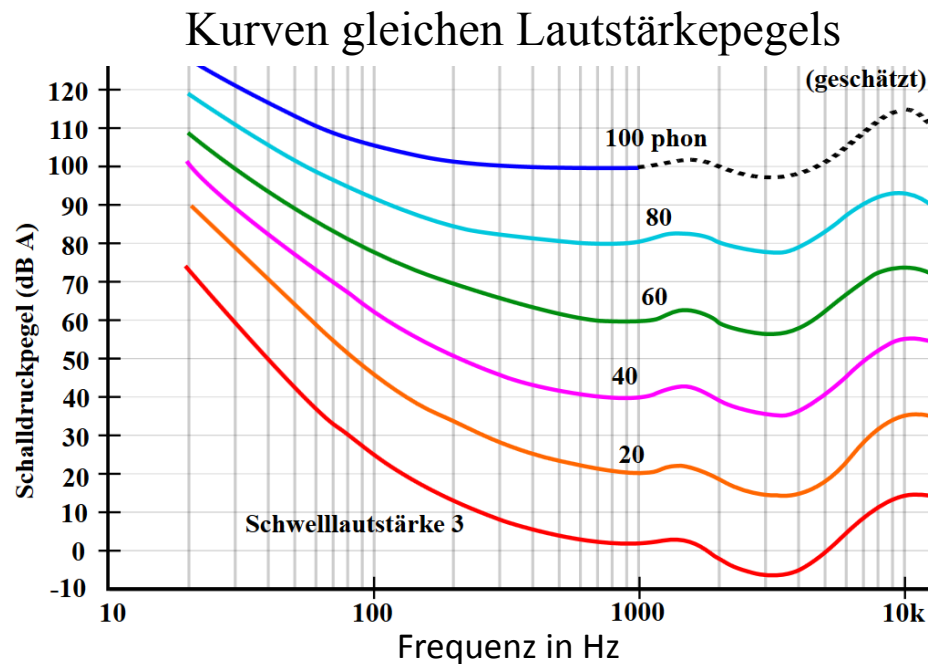
Es gibt auch Alternativen zu MFCCs

Alternative Merkmale: Perceptual Linear Prediction (PLP)

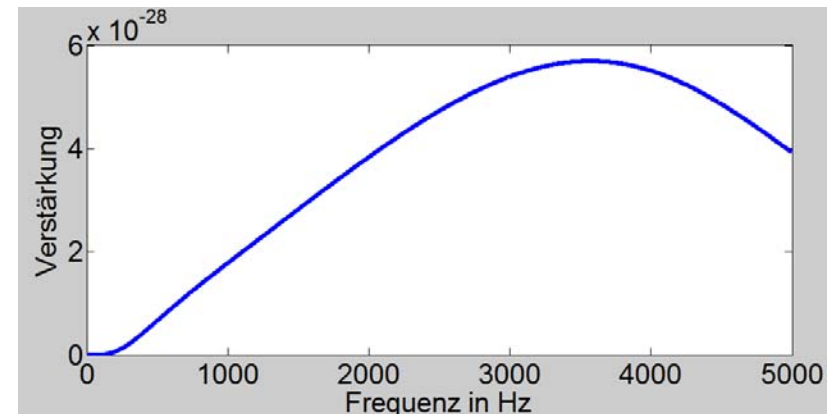
- Vorgeschlagen von Hermansky (1990)
- Ziel: *niedrigdimensionale Merkmalvektoren* für die *sprecherunabhängige Spracherkennung*. Ein Merkmalvektor sollte sprecherabhängige Information im Signal unterdrücken, die linguistische Information aber erhalten.
- Berechnungsschritte sind an der menschlichen akustischen Wahrnehmung orientiert:
 1. Spektralanalyse zur Berechnung des Kurzzeit-Leistungsspektrums
 2. gehörgerechte Verzerrung der Frequenzachse (Mel-Skala)
 3. Berücksichtigung der frequenzabhängigen Lautstärkewahrnehmung
 4. Berücksichtigung der Abhängigkeit zwischen Intensität und Lautstärke
 5. Approximation der spektralen Hüllkurve durch ein Allpol-Modell 5. Ordnung → Die Modellkoeffizienten bilden den Merkmalvektor.

Alternative Merkmale: Perceptual Linear Prediction (PLP)

Zu 3.) Berücksichtigung der frequenzabhängigen Lautstärkewahrnehmung:



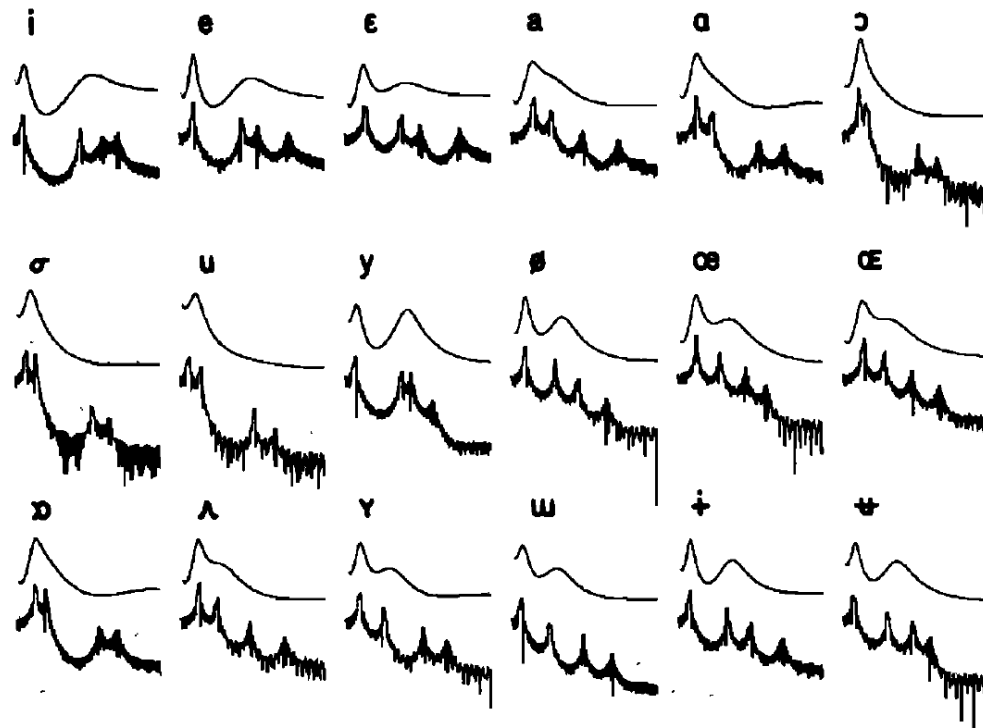
Filter zur Kompensation der Kurve gleicher Lautstärke (bei 40 dB)



Zu 4.) Berücksichtigung der Abhängigkeit zwischen Intensität I und Lautstärke L : $L(\omega) = I(\omega)^{1/3}$

Alternative Merkmale: Perceptual Linear Prediction (PLP)

Zu 5.) Approximation (Glättung) der spektralen Hüllkurve durch ein Allpol-Modell 5. Ordnung:



Bildquelle:
Hermansky (1990)

Ergebnis: Bei der sprecherunabhängigen Erkennung von Ziffern wurden mit PLP bessere Ergebnisse erzielt als mit einer klassischen LP-14-Approximation des Spektrums.

Literatur

- Euler S (2006). *Grundkurs Spracherkennung*. Vieweg-Verlag, Wiesbaden.
- Hermansky H (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4), 1738-1752.
- Stahl V (2010). *Spracherkennung*. Vorlesungsskript.