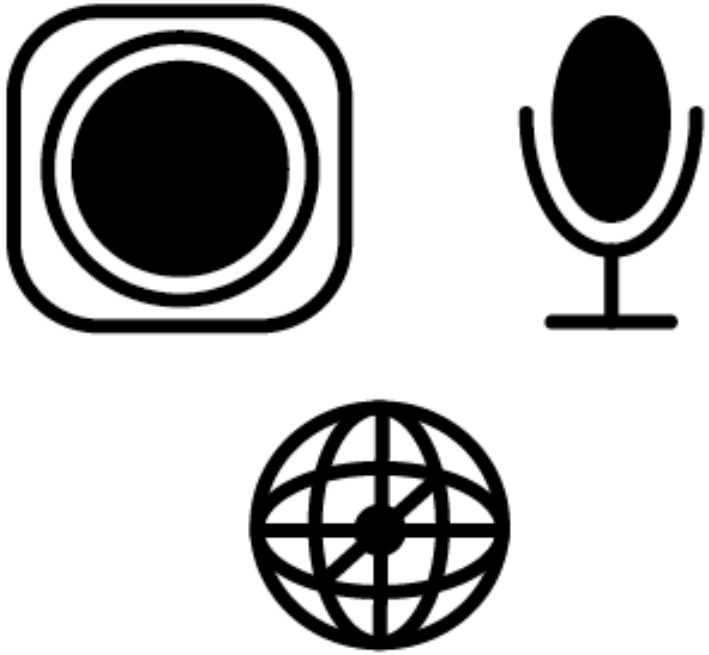


Sensordatenverarbeitung

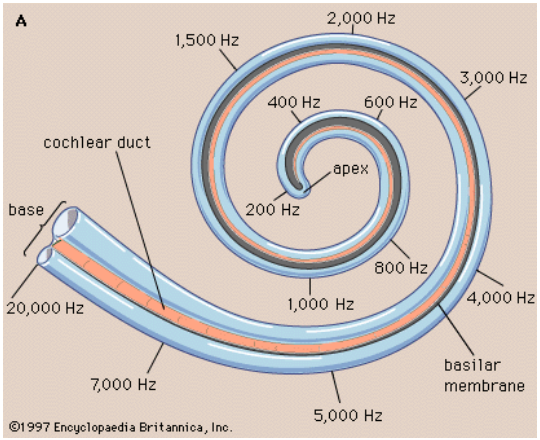
MERKMALSEXTRAKTION (10)

16.-20.12.2024

Teil c

Merkmale von Sprache

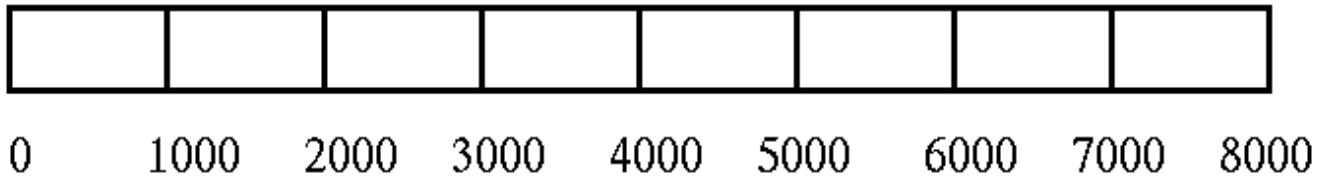
- Hier: Klassische Form der **parametrischen Repräsentation** von Sprache
- Wir erinnern uns an das menschliche Ohr
- Zerlegung in Frequenzbereich
- Nehmen wir doch spektrale Koeffizienten aus der Fourier-Transformation
- ABER: diese Koeffizienten repräsentieren auch die Mikrostruktur des Signals
- Diese hat viele Redundanzen und irreführende Informationen, wie Rauschen



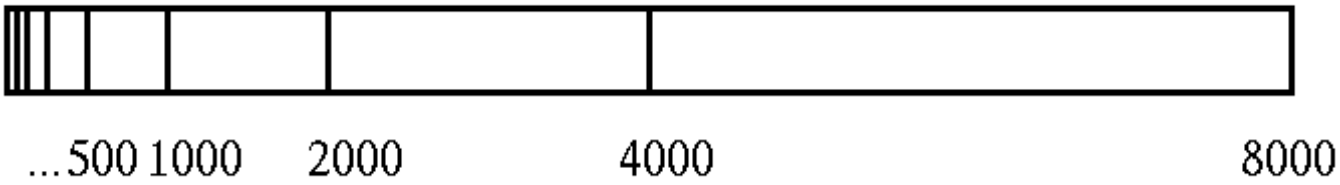
- Lösung: Filterbänke
  - Fasse Nachbarfrequenzen geeignet zusammen, bspw. durch eine (gewichtete) Mittelung benachbarter Frequenzen
  - Dies führt zu einer robusteren Repräsentation des Sprachsignals
  - Merkmalskompression erhöht die Leistungsfähigkeit
- Filterbänke werden auch im Ohr angewendet

Typen von Filterbänken

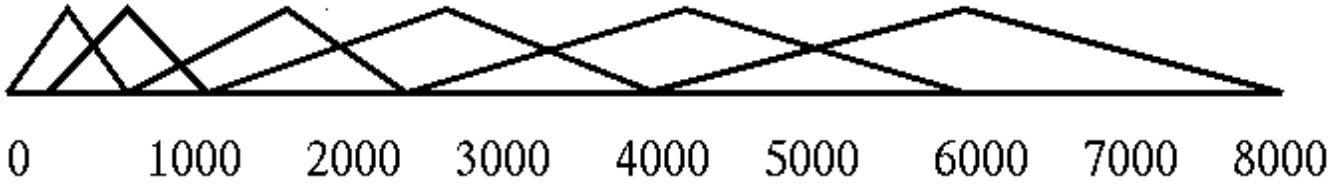
Feste Breite



Variable Breite



Überlappende Filter



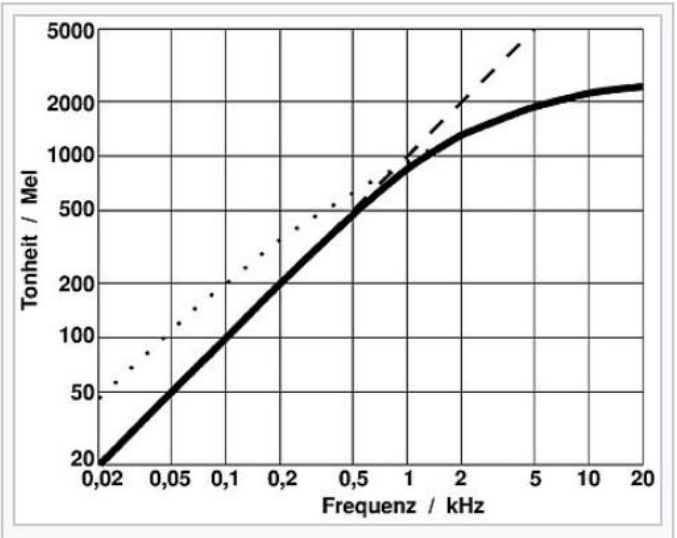
Frequenz (in Hertz)

Die Mel-Skala

- **Mel**(ody) ist die Maßeinheit für die psychoakustische Größe **Tonheit,** die die wahrgenommene Tonhöhe beschreibt [Stevens, Volkman, Newman, “A scale for the measurement of the psychological magnitude pitch”, 1937]
- Mel-Skala: ein Ton, der doppelt so hoch wahrgenommen wird, erhält den doppelten Tonheitswert
- Psychoakustische Versuche ergeben eine Tonheitsskala
  - Personen hören Töne an, geben ihre Wahrnehmung an Tonhöhen gleicher wahrgenommener Distanz vs. tätsächlicher Distanz
- Das Resultat:
  - Frequenzen <500Hz: Tonheit und Frequenz hängen linear zusammen
  - Frequenzen > 500Hz: nichtlinearer Zusammenhang, d.h. Tonintervalle werden kleiner wahrgenommen als sie tatsächlich sind

    ⇒Töne in niedrigeren Frequenzen sind wichtiger als höhere (fürs menschliche Ohr)

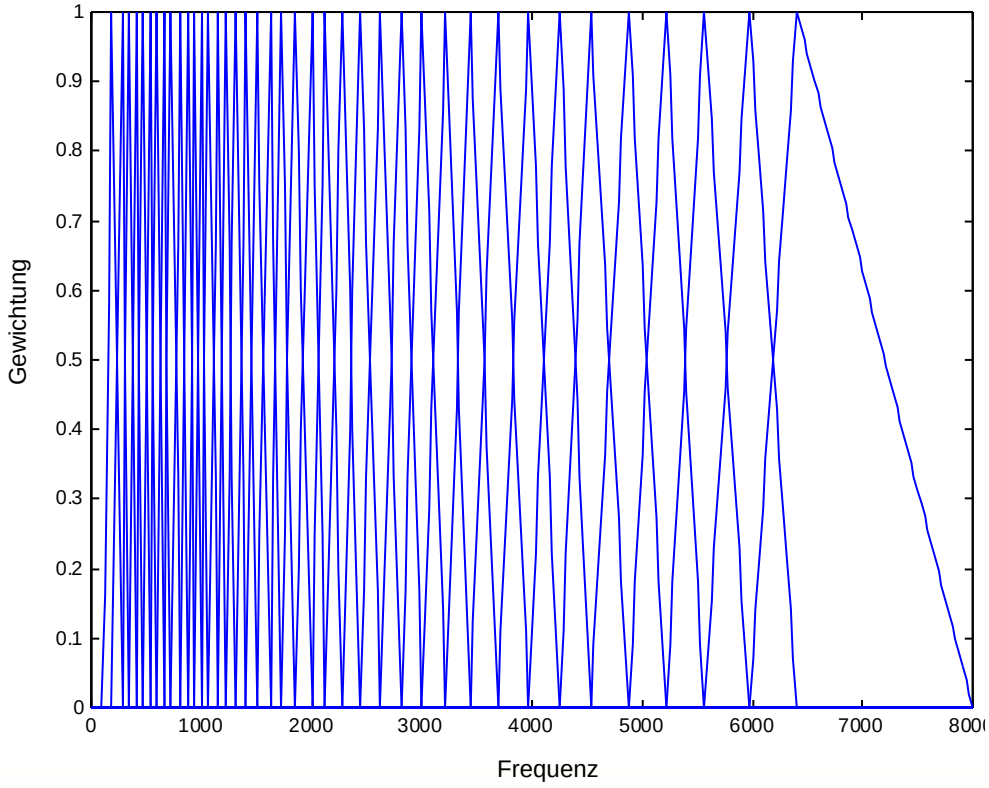
  - Der Melscale-Filter approximiert diese physiologischen Eigenschaften des menschlichen Ohrs



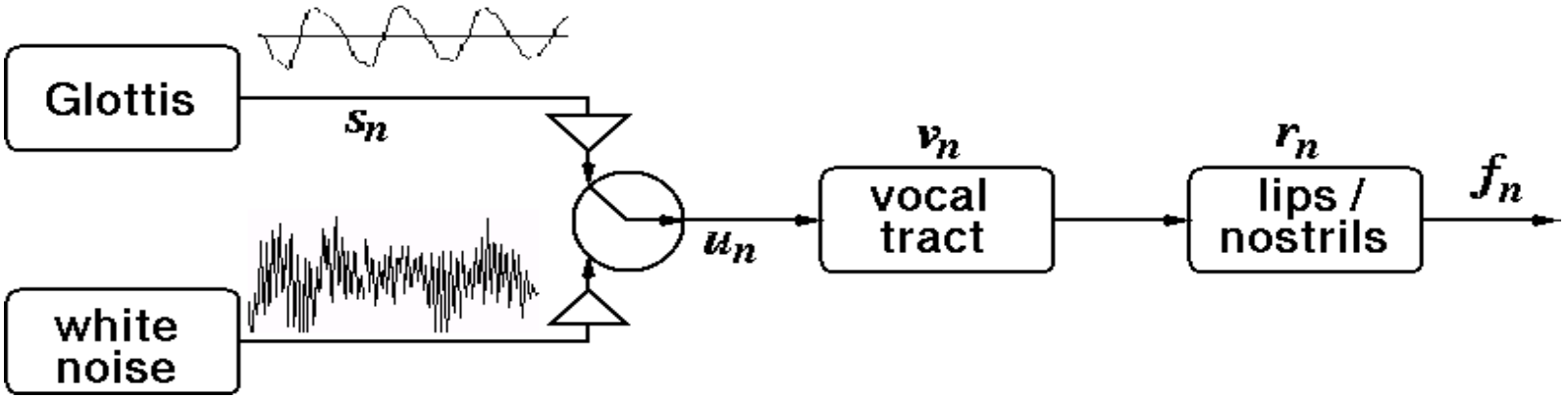
wikipedia

Melscale Filterbank

- Die resultierende **Melscale Filterbank** sieht so aus:
  - Dreieckige Gewichtungsfunktion
  - Zusammenfassung zu 25 – 40 Koeffizienten
  - Höhere Auflösung für niedrige Frequenzen

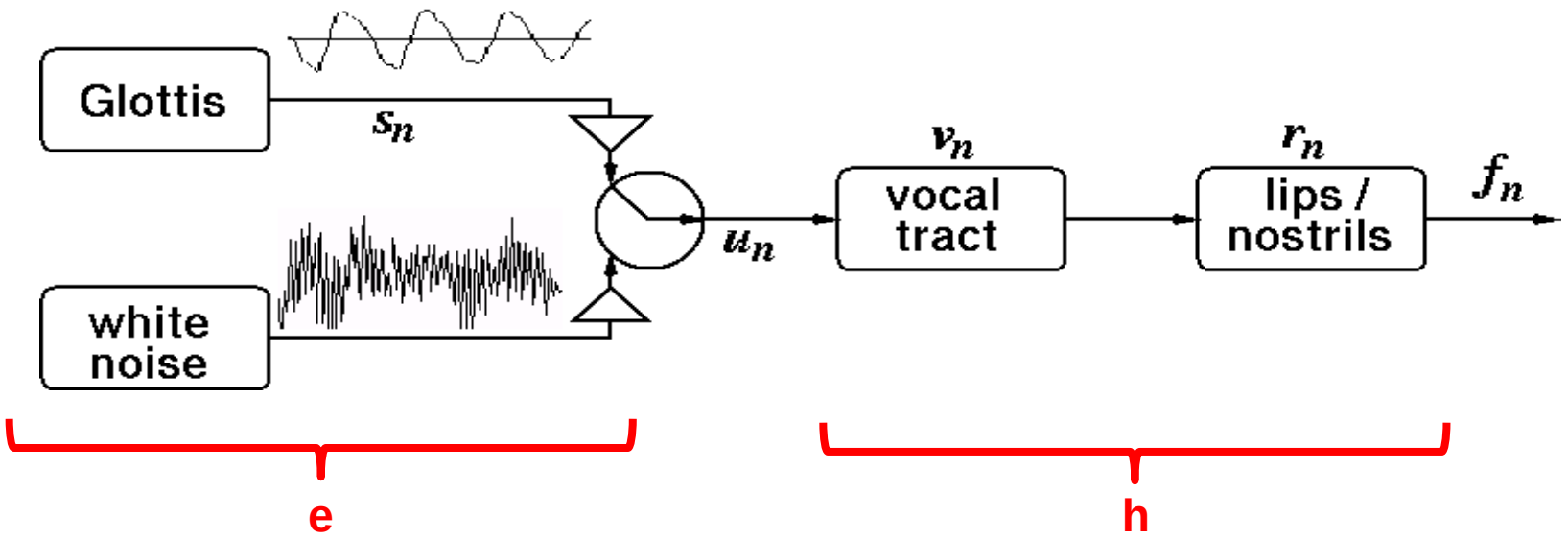


Quelle-Filter Modell von Sprache



- Laute werden erzeugt durch:
  - Vibrieren der Stimmbänder (stimmhafte Laute wie Vokale /a/, /i/, /u/, …)
  - Rauschen durch Reibung (stimmlose Laute wie /p/, /t, /k/, /r/)
  - Beides (stimmhafte Reibelaute wie /z/, /v/, …)
- Das Signal u_n wird durch den Vokaltrakt moduliert, mit Impulsantwort v_n
- Dies wird durch Lippen/Nasenloch moduliert, mit Abstrahlungsantwort r_n
- Das resultierende Signal f_n erklingt: $\boldsymbol{f_n = u_n * v_n * r_n}$

Quelle-Filter Modell von Sprache



- Die Anregungsfunktion e ist die Stimmbandanregung und/oder weißes Rauschen
- Der Filter h ist die Modulation dieser Anregung durch den Vokaltrakt
- Die resultierende Modulation f ergibt sich aus der Faltung der Anregungsfunktion e mit dem Filter h:

$$f = e * h$$

Cepstral Koeffizienten

- Quelle-Filter-Modell: $f = e * h$
- Im Frequenzbereich:

$$F(f) = F(e) \cdot F(h)$$

$$\log F(f) = \log F(e) + \log F(h)$$

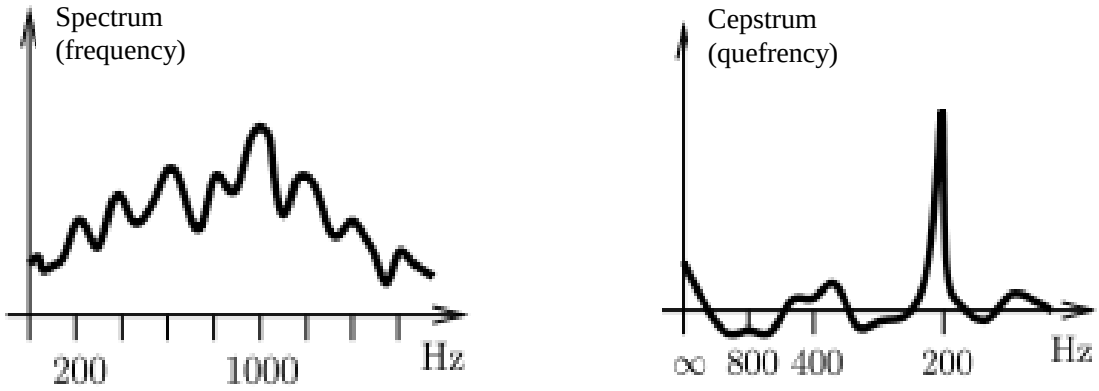
$$F^{-1}(\log F(f)) = F^{-1}(\log F(e)) + F^{-1}(\log F(h))$$

- Damit kann man Sprache entfalten in die **Summe** zweier Anteile:
  - Die niederfrequente periodische Anregungsfunktion e
  - Die Filterung des Vokaltrakts (Formanten) h
- Die Koeffizienten dieser Transformation nennt man **Cepstralkoeffizienten (**engl. cepstral coefficients) oder einfach Cepstrum
- Das Cepstrum ist die *inverse Fourier Transformation des Logarithmus der Magnitude des Spektrums*
- Anregung und Filter werden also kombiniert auf unterschiedliche Weise
  - Gefaltet im Zeitbereich
  - Multipliziert im Frequenzbereich
  - Addiert im Ceptralbereich

Cepstral Koeffizienten

- Das Cepstrum kann man interpretieren als die *Information über die Änderungsrate in den verschiedenen Frequenzbändern*
- Der Name "*cepstrum*" wurde abgeleitet aus "***spec****trum*"
- Das Cepstrum wird gemessen in "*quefrency" (von **freque**ncy)*
- Manipulationen des Cepstrums nennt man "*liftering" (von **filte**ring)*

- Beispiel: Die Tonhöhe und die Harmonischen im Spektrum (links) erscheinen als Peak im Cepstrum bei 200 Hz



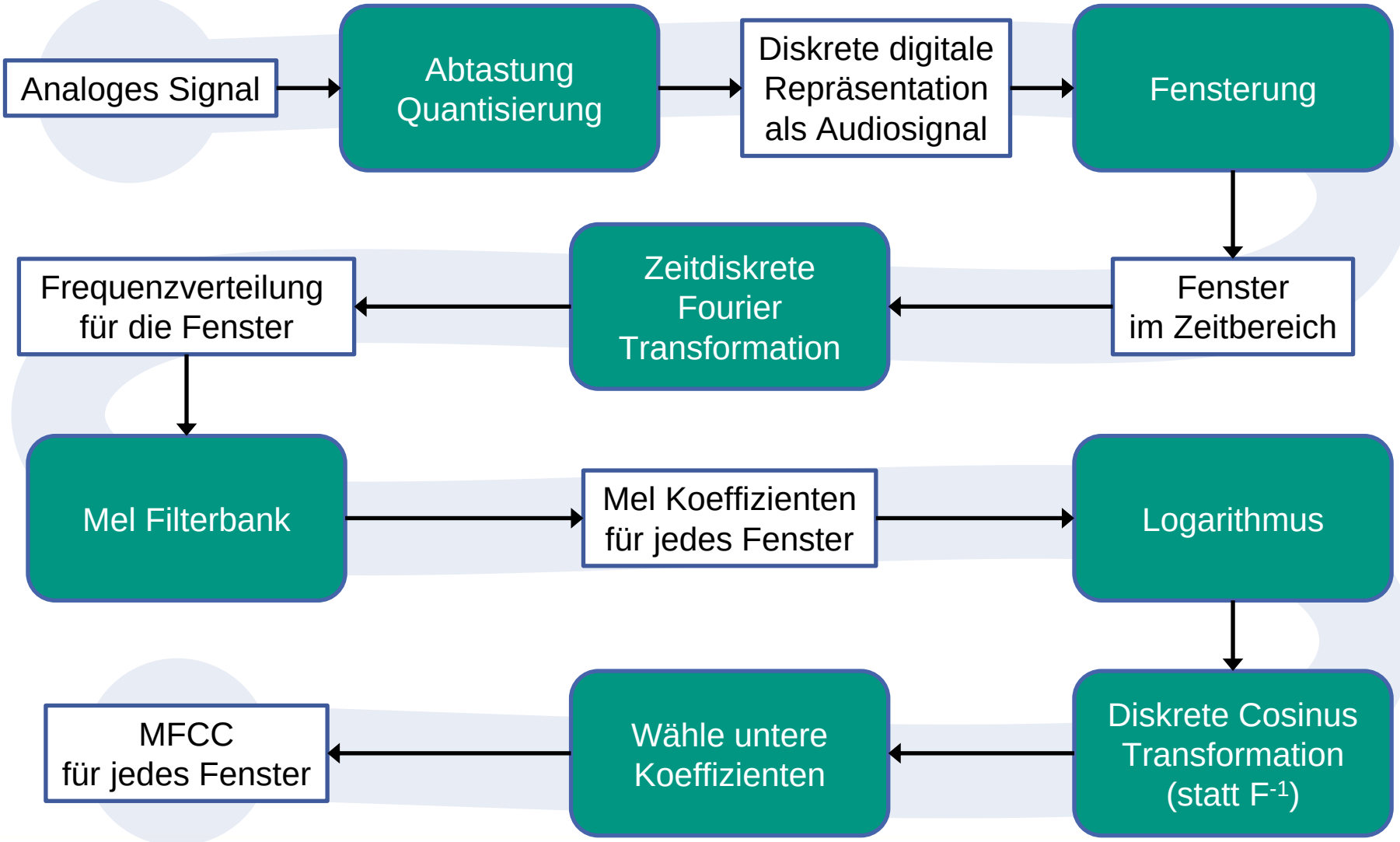
Mel-frequency Cepstral Coefficients

- Cepstral Koeffizienten haben eine Struktur
  - Die unteren Koeffizienten reflektieren die Macrostruktur des Spektrums
  - Die oberen Koeffizienten reflektieren die Microstruktur des Spektrums
  - Der 0^{th} Koeffizient reflektiert die Signal Energie

- Daher vewendet man für eine generalisierende parametrische Darstellung von Sprache die *unteren* Cepstral Koeffizienten
  - Üblich sind 13 Koeffizienten

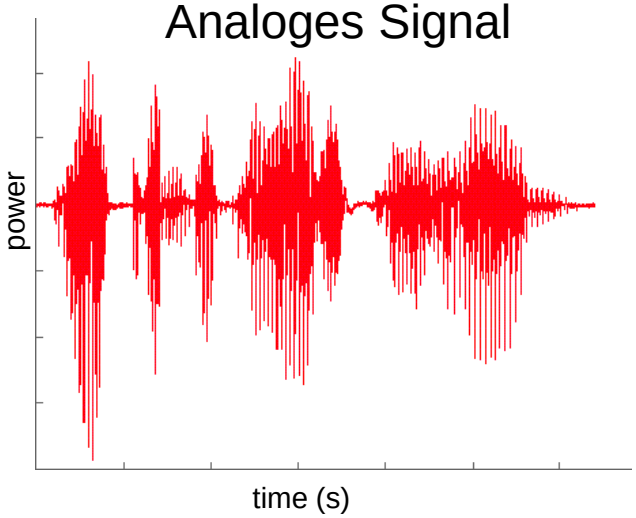
- Durch die Mel-Filterung wurde die Dimensionalität des Signals reduziert, was die Spracherkennung deutlich verbesert
- Das Resultat nennt man
  → **Mel-frequency Cepstral Coefficients (MFCC)**
  - Die Makrostruktur gibt die Eigenschaften des Vokaltraktes wieder
  - Die Mikrostruktur enthält in der Regel Rauschen, auch ist sie schwieriger zu modellieren, daher wird sie entfernt

Berechnung der MFCC

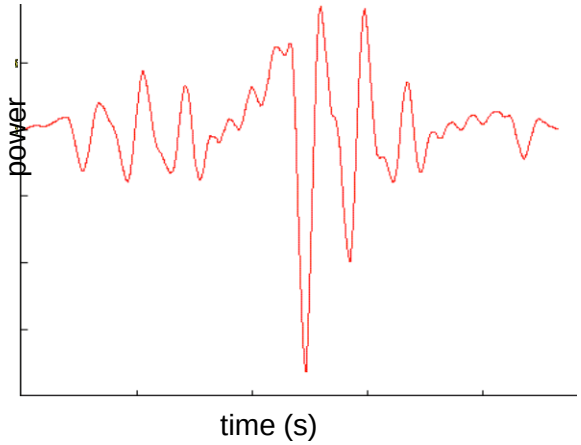


Beispiel

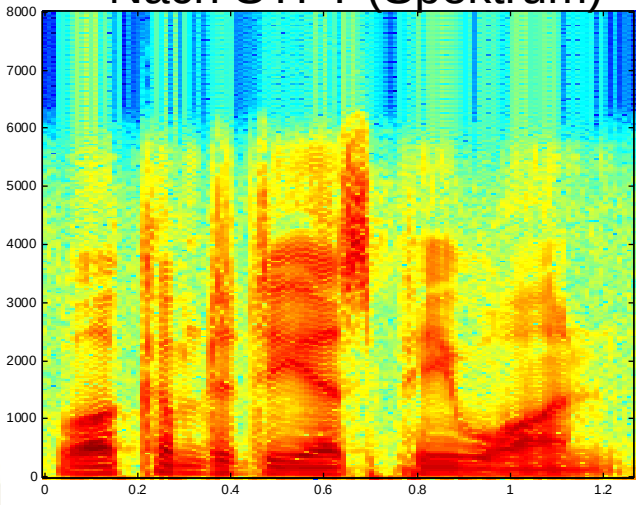
Analoges Signal



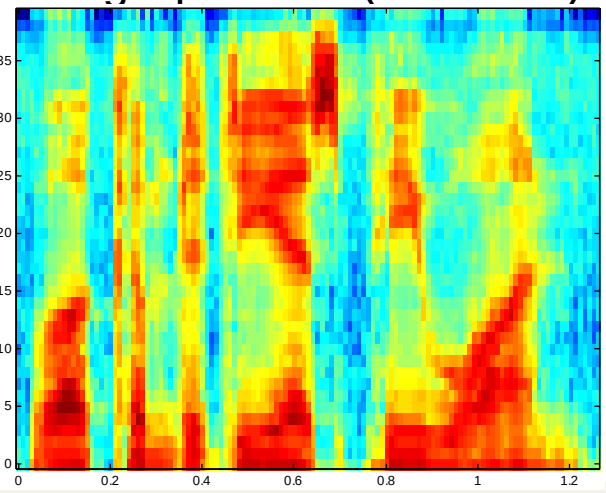
Fensterung, Vokal “u”



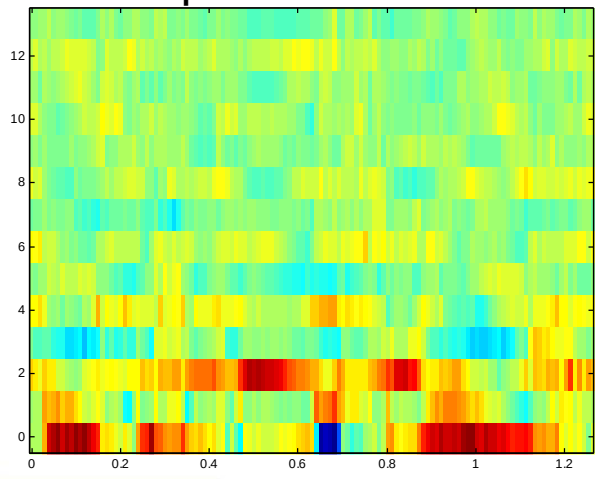
Nach STFT (Spektrum)



Log Spectrum (Melscale)

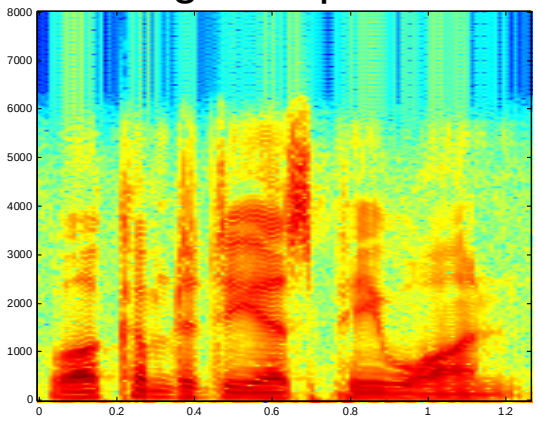


13 Cepstral Koeff.

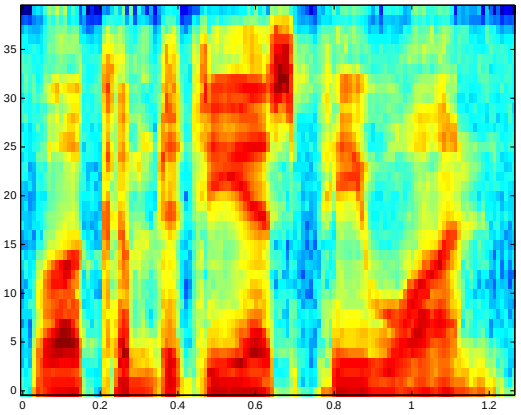


Repräsentation - Recovered

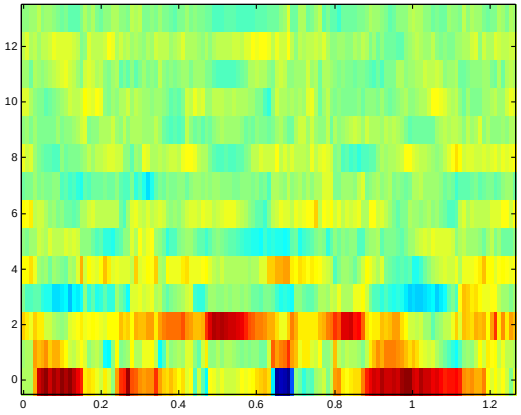
Original Sprache



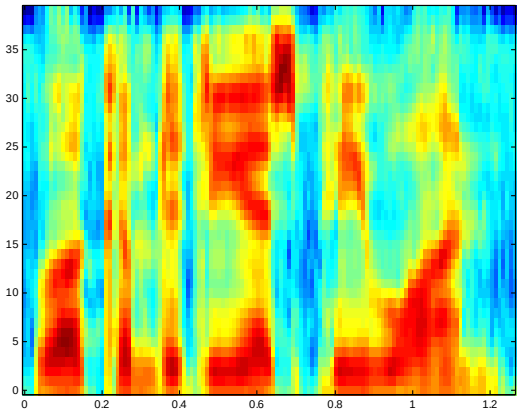
Log Mel



Cepstrum

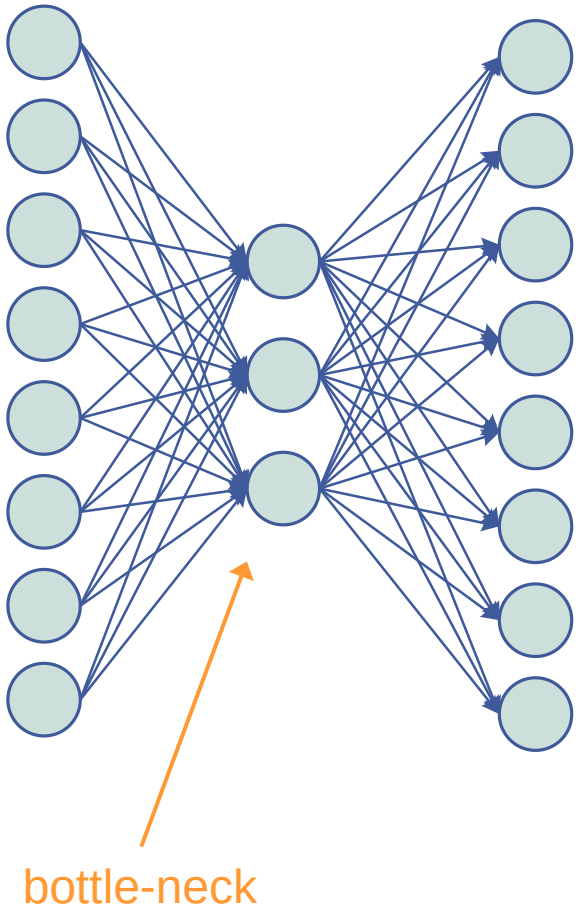


Log Mel (aus Cepstrum wiederhergestellt)



Bottle-Neck (Flaschenhals) Merkmale

- Bottle-Neck Merkmale werden durch Neurale Netze erstellt Multilayer Perceptron (MLP)
- Der Flaschenhals ist eine versteckte Schicht, die viel kleiner ist als andere Schichten des Netzes
- Eingabe-Information (Sprachrepräsentation), die relevant ist zur Klassifizierung an der Ausgabeschicht (Klassen) muss durch den Flaschenhals, d.h.
- Nur wenige Knoten müssen die wichtigste Information speichern
- Wenn das MLP mit den richtigen Targets trainiert wird, enthält die Flaschenhals-Schicht eine gute Kodierung der Eingabeinformation (im Sinne der Klassifikation der Targets)

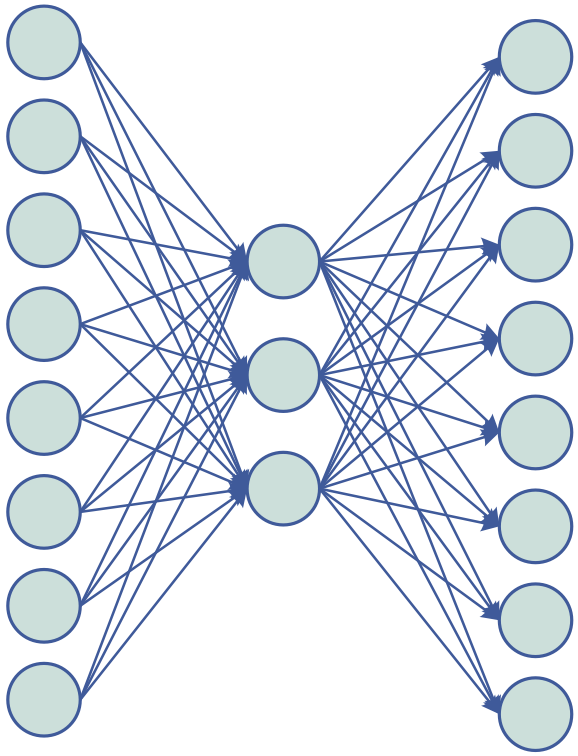


Beispiel: Auto-Encoder

- Auto-Encoder

  Ein neuronales Netz, das lernt, die Eingabe an der Ausgabeschicht zu reproduzieren

- Beispiel
  - Topologie wie rechts abgebildet 8-3-8
  - Trainingsbeispiele und Klassen:
    Die Ziffern 0-7
  - Encoding der Ziffer i
    - Wert am Knoten i 1
    - Wert an anderen Knoten 0
    - z.B. Ziffer 5 = 00000100
  - Resultat:
    - Die drei Knoten der verborgenen Schicht lernen die Darstellung der Ziffern im *Binärkode – die kompakteste Darstellung von Ziffern*
    - D.h. Ziffer 5 = 101

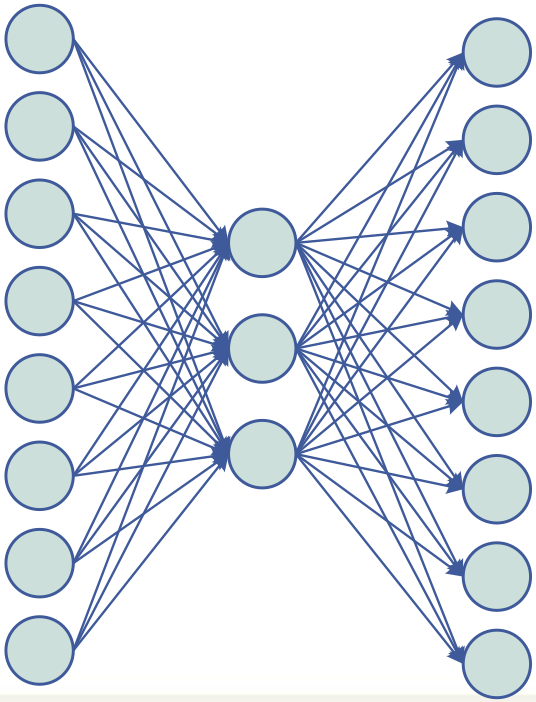


Bottle-Neck Merkmale für Sprache

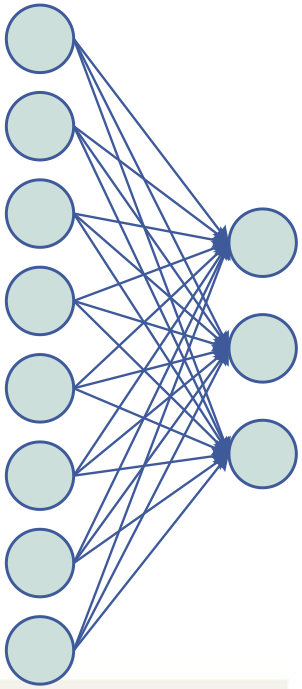
- Eingabeschicht: MFCC Merkmale
- Targets für Ausgabeschicht:
  - Was wird in Spracherkennung klassifiziert: Laute

→Nutze Laute (Phone/Phoneme) als Targets

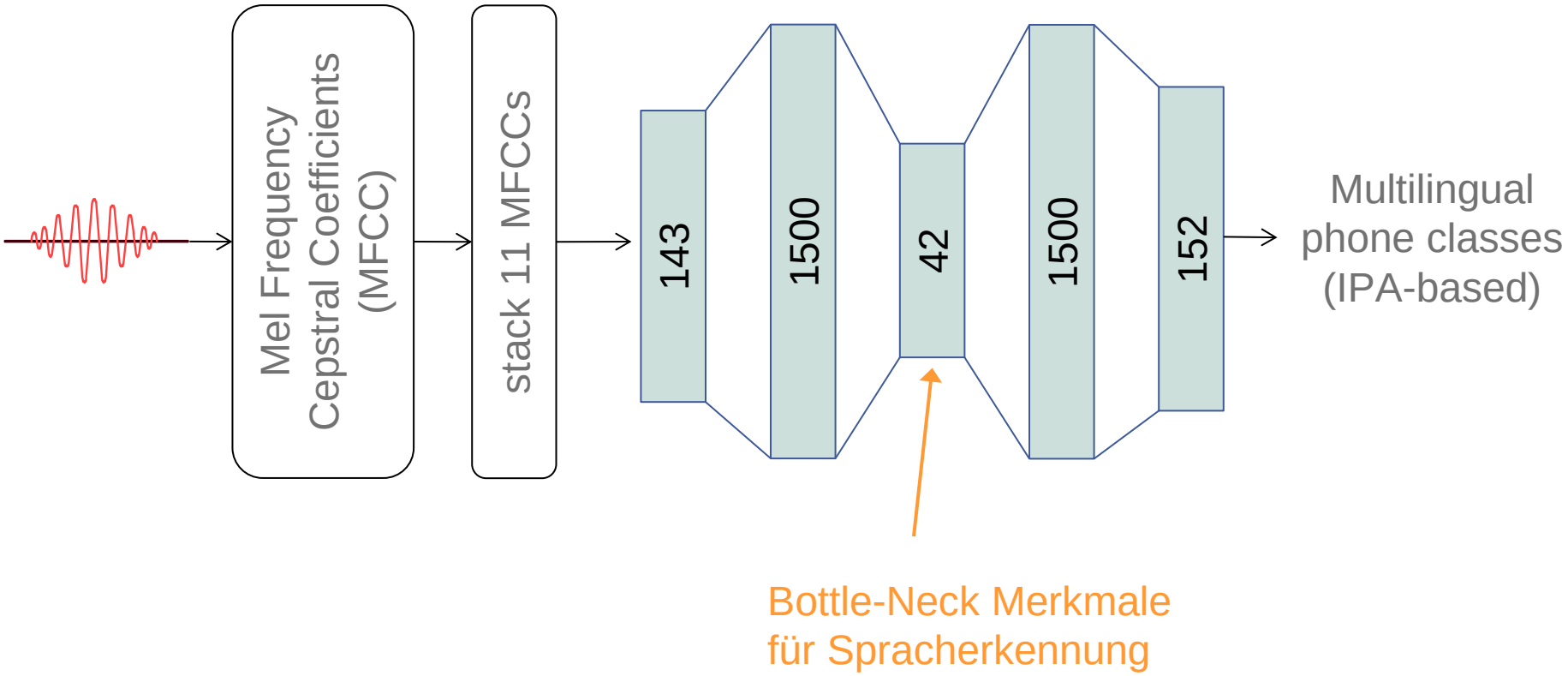
- Schritt 1 Trainiere NN



Schritt 2: Nutze Bottle-neck als Merkmale



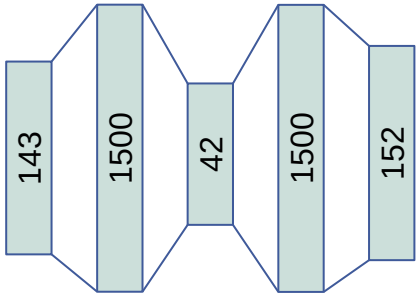
Ein Beispiel aus der Praxis



Herausforderung: Daten

- Ein Neuronales Netzwerk hat VIELE Parameter

Ein Gewicht pro Kante	568,500
Ein Bias pro Knoten	3,194
Total # Parameter	571,694



- Zum Lernen dieser Parameter benötigt ma viele Daten
  - Für viele Anwendungen hat man nicht so viele

- Neuronale Netze sind “black box”
  - Große Verbesserungen in der Spracherkennung aber:
  - Was lernt das Netz?
  - Wie repariert man Fehlklassifikationen?

Zusammenfassung

- Was sind Merkmale
  - Definition
- Wozu braucht man die
  - Klassifikation, Beispiele
- Welche Merkmale sind gute Merkmale
  - Was sind gute Merkmale
    - Signal, Anwendung
    - Unterscheidend, Kompakt, Robust
    - Herausforderungen: Fluch der Dimensionen
- Wie extrahiert man Merkmale
  - Short-time Fourier: zeitlicher Verlauf
  - Merkmalsextraktion für Sprache
    - Quelle-Filter-Modell, Melskala, Cepstral Koeffizienten
  - Bottle-Neck Merkmale, Autoencoder, Tiefe Neuronale Netze
- Verfahren zur Merkmalsselektion und Merkmalsreduktion
  - In SdV nicht ausreichend abgedeckt
  - Siehe bswp. Grundlagen des Maschinellen Lernens