0.1 Eigenschaften gesprochener Sprache

Sprachsignale haben u.a. folgende Eigenschaften:

- stimmhafte vokalische Abschnitte sind gekennzeichnet durch Frequenzbereiche, die sich im Spektrogramm als Bänder über ca 50-250ms Dauer erkennen lassen.
- Ihnen entsprechen sog. Formanten, das sind Resonanzbereiche des Vokaltrakts.
- Vokale sind durch 2-3 Formanten charakterisierbar (bezeichnet als $F_1, ..., F_3$). Dabei variieren alle Formanten über die Zeit, über mehrere Sprachsamples einer Person, sowie zwischen Personen. Ferner können Formanten in Diphthongen (Doppelvokalen) in einander übergehen.
- stimmlose Abschnitte sind durch geräuschhafte Anteile im Sprachsignal erkennbar. Diese sind z.B. bei Plosivlauten deutlich kürzer als bei Zischlauten und erreichen meist nicht die Länge von Vokalen.
- Konsonanten sind weitaus mehr als Vokale anhand des Kontextes erkennbar. Bspsw. hängt die Unterscheidung b vs p vom zeitlichen Abstand zwischen Plosiv-Geräusch und Einsatz der Intonation ab.
- Einige Konsonanten werden anhand des Formantverlaufs umgebender Vokale unterschieden.
- Grundfrequenz der Stimmbänder variabel durch Intonation und abhängig von Sprechern (Kinder, Frauen, Männer). Die Grundfrequenz wird als nullter Formant F_0 bezeichnet. Sie tritt in stimmhaften vokalischen oder konsonantischen Segmenten des Sprachsignals auf.
- Koartikulation – in gesprochener Sprache werden akustische Kennzeichen phonetischer Elemente verschliffen; Sonderfall *vocalic reduction* – Angleichung von Formanten aufeinander folgender Vokale in der Koartikulation
- Prosodie bezeichnet die Variation der Grundfrequenz der Stimmbänder im Artikulationsverlauf.
- Lombardisierung – Erhebung der Stimme und Veränderung von Formanten abhängig von Umgebungsgeräuschen.

0.2 Physiologische Grundlagen frequenzselektiver Wahrnehmung

(Wir überspringen in dieser Übersicht die bildliche Darstellung und Erläuterung der physiologischen Komponenten des Gehörs.)

- Im Innenohr wird das Schallsignal in eine Wanderwelle auf der Basilarmembran transformiert. Dabei entsteht eine maximale Amplitude abhängig von der Reizfrequenz an einem bestimmten Ort der Basilarmembran (Ortsprinzip oder Tonotopie; entdeckt von Bekesy). Für hohe Frequenzen liegt das Amplitudenmaximum nahe am Mittelohr (Steigbügel) und verschiebt sich dann mit fallender Anregungsfrequenz hin zur Spitze der Cochlea (dem Helikotrema).

- Die Schwingung der Basilarmembran löst eine Abbiegung der Sinneshärchen an Rezeptorzellen auf dem Corti-Organ aus. Dieses befindet sich in der $K+$ reichen Endolymphe in der scala media.

- Die Abbiegung löst einen Transduktionsprozeß aus, der zur Öffnung von Ionenkanälen für die $K+$-Ionen und damit bei entspr. Reizung zur Depolarisation der Sinneszellen mit Erzeugung eines elektrischen Impulses führt.

- Durch die Abbiegung (Deflektion) der inneren Haarzellen werden elektrische Signal produziert, die, entsprechend dem Ortsprinzip, einen afferenten Transmitter an die anliegenden afferenten Fasern des Hörnervs weitergeben. (afferent := zum Gehirn oder einer zentralen Verarbeitung hinführend, sensorische Nervenbahnen; efferent – motorische Nervenbahnen)

- Die Deflektion der äußeren Haarzellen ist Bestandteil eines Regelungsmechanismus, der als *kochleäre Verstärkung* bezeichnet wird. Die Äußeren Haarzellen schwingen nahe der Amplitudenmaxima in der charakteristischen Frequenz mit und lösen so eine bis zu 1000 fache Verstärkung der Membranauslenkung aus. (In die äußeren Haarzellen führen überwiegend efferente Nervenbahnen. Dies läßt darauf schließen, daß es einen physiologischen Mechanismus zur Regelung des Ausmaßes der cochleären Verstärkung gibt.) Dies führt zugleich zu einer weit stärkeren Orts-Konzentration der Auslenkung, als es durch den Bekesy-Mechanismus erklärbar wäre.

- Einzelne Fasern des Gehörsnerven sind daher frequenzselektiv reizbar. Detektions- bzw. Diskriminierungsschwellen (vor Hintergrundrauschen) liegen nahe einer jeweiligen charaktistischen Frequenz um bis zu 20dB niedriger. Typischerweise sind diese tuning curves zu den hohen Frequenzen spitzer als zu den niedrigen.

0.3 Exkurs: Maße für Intensität

Wir betrachten Schall als periodische Wanderwelle durch ein Medium, üblicherweise Luft. Ein einfaches Maß für die Intensität ist der sog. RMS-Wert, der die Wurzel aus dem Integral über eine Periode der Schwingungsfunktion darstellt. Für eine Sinusschwingung mit Amplitude C ist $x_{RMS} = C/\sqrt{2}$.

In der Realität erzeugt der Druck der Schallwelle Variationen in der Geschwindigkeit der Brownschen Molekularbewegung im Medium. Dies charakterisieren wir durch einen RMS-Wert für die Geschwindigkeit u abhängig vom RMS-Wert für den Luftdruck unter Einfluß einer ebenen Wanderwelle (Schalldruck):

$$u = \frac{p}{z}[\frac{m}{s}]$$

Dabei steht z für die spezifische akustische Impedanz des Mediums. Es gilt

$$z = \rho\nu[\frac{kg}{m^2 s}]$$

wobei $\nu[\frac{m}{s}]$ die Schallgeschwindigkeit im Medium, $\rho[\frac{kg}{m^3}]$ die Dichte des Mediums beschreibt. Im Fall der Luft gilt, daß bei der Hörschwelle des Menschen bereits eine systematische Oszillation von Luftteilchen hörbar wird, deren RMS-Geschindigkeit gerade ein 10-Milliardstel der durchschnittlichen Geschwindigkeit der thermischen Molekularbewegung beträgt.

Ähnlich wie eine Wasserwelle trägt auch eine Schallwelle Energie, da sie Kraft auf die Teilchen des Mediums ausübt. Die char. Größe hierfür ist die Leistung (*power*, Stichwort Arbeit pro Zeiteinheit), ausgedrückt in der Einheit Watt, entspr. Joule pro Sek.

Die Intensität oder Leistung (*power*) I einer Wanderwelle wird entspr. in Watt pro m^2 ausgedrückt. Es gilt

$$I = pu = \frac{p^2}{z}[\frac{W}{m^2}]$$

Die menschliche Hörschwelle liegt nominal bei $I_0 = 10^{-12}[\frac{W}{m^2}]$.

Akustische Lautstärke *acoustical strength* (zu unterscheiden von wahrgenommener Lautheit *loudness*, s. unten) wird durch eine Verhältnisskala logarithmisch abhängig von der Intensität ausgedrückt:

$$L_2 - L_1 = 10\log(\frac{I_2}{I_1}) = 20\log(\frac{p_2}{p_1})[dB]$$

(wegen der Abhängigkeit der Intensität vom Quadrat des Schalldrucks p)

Nutzt man die menschliche Hörschwelle I_0 als Referenzintensität, erhält man die dB SPL (*sound pressure level*) Skala $L = 10\log(\frac{I}{I_0})$.

Wie bestimmt man die Intensität bei einem aperiodischen Schalldruck, wie er etwa durch Rauschen entsteht? Auf eine nach Einführung der Fourieranalyse zu besprechende Weise gehen wir von einer Intensitätsdichtefunktion (*power density*) aus, die uns abhängig von der Frequenz f eine Intensität $D(f)$ liefert. Bei weißem Rauschen ist diese Dichtefunktion über den hörbaren Bereich konstant. Bei einem Rauschband haben wir in einer Bandbreite Δf ein konstantes $D(f) = D_c$, so ist einfach

$$I = D_c\Delta(f)$$

Dies nutzen wir, um das sogenannte Spektrum-Level eines Rauschbandes zu definieren:

$$L = 10\log(\frac{I}{I_0}) = 10\log(\frac{D_c}{10^{-12}}) + 10\log\Delta f$$

D.h. die Lautstärke hängt additiv von $\log D_c$ und dem Log. der Bandbreite ab. Den Ausdruck

$$N_0 = 10\log(\frac{D_c}{10^{-12}})$$

bezeichnet man daher als Spektrum-Level. (N_0 ist log eines dimensionslosen Terms und daher in der Skaleneinheit dB zu verstehen.)

Ein nicht exakt rechteckiges Rauschband kann man durch ein bzgl. der Intensität gleiches rechteckiges Rauschband annähern. Die entsprechende Rechteck-Bandbreite heißt auch *equivalent rectangular bandwidth*.

0.4 Auditive Filter

- Wirkung der frequenzselektiven Wahrnehmung wird durch auditive Filter modelliert

- Filter := Komponente, die selektiv Frequenzbereiche in der Intensität anhebt oder reduziert
- auditive Filter werden beim Menschen durch Detektion (Erkennung) oder Diskrimination (Heraushören eines Signals gegen Rauschen) experimentell untersucht.

1. Band-widening experiment (Maskierungseffekt bei simultanem Reiz + Rauschen)
2. Mehrfacher Filter (Frequenzselektive Wahrnehmung als 'Filterbank')
3. Nicht-simultane Maskierung [Rauschen und Signal nicht simultan]

0.4.1 Kritische Frequenzbänder

- das Fletcher Experiment [band-widening]: Detektion eines stufenweise leiser werdenden Tons gegen Hintergrundrauschen verschiedener Bandbreite. (DEMO)
- Voraussetzung: Rauschen hat über die verschiedenen Bandbreiten jeweils konstanten Spektrallevel (N_0 bzw. entsprechende Leistungsdichte D_{c0}): d.h. Rauschen klingt bei schmaleren Bändern leise, Leistungsdichte je Frequenzbereich bleibt aber gleich.
- Nur Vergrößerung der Bandbreite innerhalb eines engen Frequenzbandes im Rauschen trägt zur Verringerung der Detektion des Signaltones bei (-> krit. Band)
- innerhalb des krit. Bandes gilt: Je breitbandiger das Rauschen, desto geringer die Detektion
- Ab einer bestimmten Rauschbandbreite wird Detektion des Signals nicht mehr zusätzlich beeinträchtigt [kritische Bandbreite].

Frequenzselektive auditive Wahrnehmung:

- Frequenzselektivität: selektive Detektion nach Frequenzbereichen.
- Gehör verhält sich wie eine Bank von Filtern; dabei wird lt. Fletcher für jedes (Sinus-)signal genau 1 Filter benutzt -> auditive Filter

- Experiment liefert -> Menge von kritischen Bändern, die die Frequenzempfindlichkeit des Gehörs kennzeichnen
- heutiger Erkenntnisstand: auditive Filter überlappen, d.h. wir hören tatsächlich quasi auf mehreren Kanälen. Ferner sind die Filterkennlinien nicht rechteckförmig.
- Leistungsspektrummodell der auditiven Filter [Fletcher]

  N_0**:** spectrum level, dazu gehörend konstante Leistungsdichte D_{c0}

  CB**:** kritische Bandbreite

  P_s**:** Leistung, bei der Signal s 'verdeckt' wird [je nach exp. Definition des betreffenden jnd]

- Annahme Fletchers: Der Quotient aus Leistung des Tones und des Rauschens im Bereich des kritischen Bandes ist bei einsetzender Maskierung konstant, d.h., weder vom gewählten Frequenzbereich noch dem eingesetzten Spektrallevel abhängig. D.h., wenn das Rauschen den Ton gerade verdeckt, dann gilt die Bez.

$$\frac{P_s}{CB\, D_{c0}} = K$$

  Dabei ist $CB\, D_{c0}$ die Leistung des Rauschens, die auf den auditiven Filter fällt.

- Verfahren:
  - N_0 variieren bzw. experimentell fixieren
  - P_s beobachten -> CB(krit. Bandbreite schätzen)
- Annahme Fletcher: $K = 1$ über den ganzen hörbaren Frequenzbereich. Heute eher K ca. 0.4 d.h. $CB = 0.4\frac{P_s}{N_0}$

Darstellung der Frequenzwahrnehmung durch eine Filterbank, wobei die Frequenzantwort der Filter den Eigenschaften der kritischen Bänder entsprechen sollen.

1. Annäherung durch Rechteckfilter: ERB: equivalent rectangular bandwith

2. Empirische Erhebung der Charakteristiken der Gehörsfilter.

   (a) auf neuronaler Ebene Isointensitätskurve tuning curves over frequency

   (b) durch Maskierungsaudiogramme

   (a) s. Folien aus Moore

   (b) Für Töne verschiedener Frequenz und Intensität wird jeweils die zur Maskierung nötige Intensität gemessen.

0.4.2 Auditive Filterbank

Die auditive Filterbank, d.h. ein Array von frequenzselektiven Filtern in der auditiven Signalverarbeitung läßt sich deutlich am Einfluß kritischer Bänder auf die subjektive Lautheitssummation zeigen –

- Innerhalb eines kritischen Bandes trägt eine Verbreiterung des Rauschbandes weniger zur Lautheit bei, als wenn das Bandrauschen über mehrere kritische Bänder geht. Es gibt eine wahrnehmungsbedingte *Lautheitssummation über kritische Bänder*.

- Demo: Illusion größerer Lautstärke wenn kurzes Rauschen bei gleicher Gesamtenergie breitbandiger wird, d.h. über mehrere auditive Filter wahrgenommen wird.

0.4.3 Auditory Streaming

Ebenfalls beeinflußt die krit. Bandbreite die Wahrnehmung in der Frequenz benachbarter Töne als zusammengehörig (DEMO Audit. Streaming). Der Streaming-Effekt ist allerdings zusätzlich durch temporale und rhythmische Charakteristika bestimmt.

0.4.4 Nutzung kritischer Bänder bei der Audiokompression

Maskierungs- und Verdeckungseffekte innerhalb kritischer Bänder werden bei der Kompression von Audiosignalen in mp3 etc. genutzt (DEMO).

Dabei besonders wichtig – wegen Überlappung der Auslenkungsmaxima auf der Basilarmembran maskieren tiefe Frequenzen höhere stärker als umgekehrt. [Demo: 'upward spread of masking']

0.4.5 Vorwärts- und Rückwärtsmaskierung

Maskierungseffekte treten auch bei zeitlicher Versetzung von Signal und Maskierer auf. Diese tlw. erst ansatzweise verstandenen Phänomene spielen eine wichtige Rolle bei der Unterscheidung von Konsonanten im Sprachsignal.

In beiden Fällen (Vorwärts- und Rückwärtsmaskierung) wird ein breitbandiges Rauschen eingesetzt, vor oder kurz nach dem ein kurzer Sinuston innerhalb der kritischen Bandbreite gespielt wird. Der Maskierungseffekt wird für mehrere zeitliche Abstände jeweils anhand einer Serie von Wiederholungen ermittelt, bei denen die Lautstärke des Sinustons jeweils um 4-6 dB abgesenkt wird.

Vorwärts-Maskierung: Das Rauschen maskiert einen nachfolgenden Ton. Dies wird auf die Refraktärphase der Sinneszellen und Neuronen in der auditiven Bahn zurückgeführt.

Rückwärts-Maskierung: Das Rauschen maskiert einen vorausgehenden (!) Ton. Dies wird auf eine Integrationsleistung in der kognitiven Verarbeitung zurückgeführt. Das Phänomen weist Ähnlichkeit zur Unterscheidung von Konsonanten aufgrund des ztl. Abstandes zw. Geräusch und Vokal auf.

0.4.6 Nutzung für spracherkennende Systeme

Die auditiven Filter erklären unsere Fähigkeiten, Sprache aus frequenzlokalen Störungen quasi herauszuhören. Diese äußern sich z.B. im Cocktail-Party-Effekt (Heraushören eines Sprechers aus vielen Sprechern). Heute werden dem auditiven System nachgebildete Filterbänke in front ends von Spracherkennern genutzt.