Hearing and Speech Processing

Biomedical Engineering Year 3 Notes

2020-2021

1 Introduction to the auditory system

1.1 Sound: some definitions

- **Intensity:** Amplitude of the pressure wave, measured in dB SPL (decibel sound-pressure level)
- **Loudness:** Subjective, perception of sound strength, measured in phon
- **Number of phon of a sound:** The dB SPL of a sound at a frequency of 1 kHz that sounds just as loud.
  - 0 dB SPL at 1 kHz is hearing threshold
  - 0 phon: limit of perception
  - sound with negative phon: inaudible
  - sound with positive phon: audible

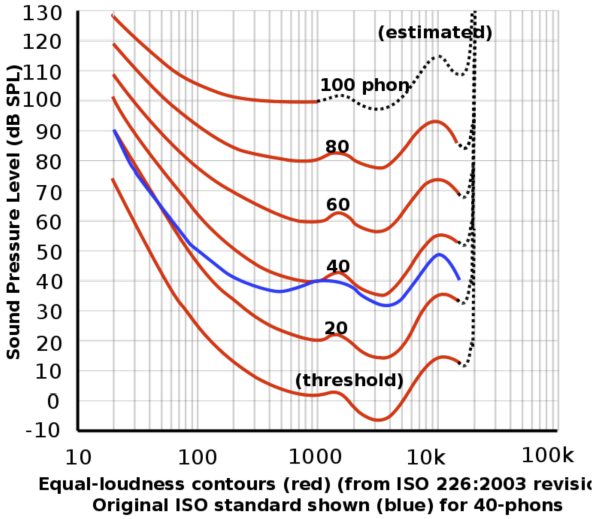

Figure 1: Equal loudness contours

1.2 An Intro to Sensitivity, Selectivity and Invariance in the auditory system (AS)

- AS detects atomic-scale vibrations
- AS resolves frequencies at around 0.5% microsecond time resolution
- Auditory neurons are excited only by specific combination of features
- We can recognize distorted and noisy speech
- Unlike fingerprints, most natural patterns are not fixed, but are variable. Recognition of variable patterns requires not only selectivity, but also invariance. This is why pattern recognition is a hard problem.
- Sensitivity, Selectivity, and Invariance are key for any sensory system (natural or artificial).

1.3 Sound: some physics

Sound is a pressure wave that propagates at a speed V.

- In air, $V \approx 343$ ms^{-1}
- In water, $V \approx 1484$ ms^{-1}
- In rock, $V \approx 5000 - 9000$ ms^{-1}

All frequencies reach the ear at the same time. The ear decomposes a pressure wave into frequency channels. A set of frequency-selective detectors is required for this decomposition to work.

2 Outer, middle and inner ear

2.1 Outer ear

It establishes spectral cues used for vertical (elevation) sound localisation.

- The human auricle (outer ear, pinna) has a peculiar shape that plays an important role in sound localisation.
- Collects sound energy
- Filters sound spectrum as a function of direction (elevation), establishing spectral cues used for vertical (elevation) sound localisation.
  - This relies on the shape on the ear
  - It can be tested by inserting a mold into the ear to modify the transfer function and see how this affects sound localisation
  - After a while with the mold, sound localisation can be relearned and re-calibrated

2.2 Middle ear

The middle ear provides impedance matching between air and inner-ear fluid. Tympanic ears evolved independently in at least three groups of terrestrial vertebrates - this shows that there was a high evolutionary pressure for land animals to have them. Tympanic ears provide sharper hearing.

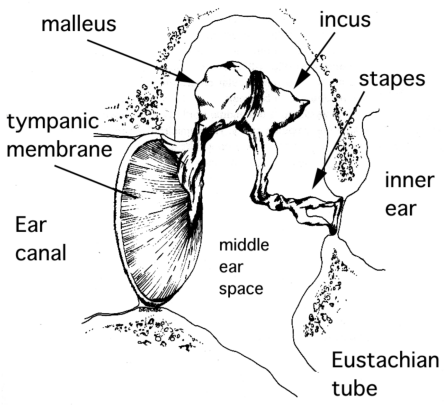

Figure 2: Middle Ear

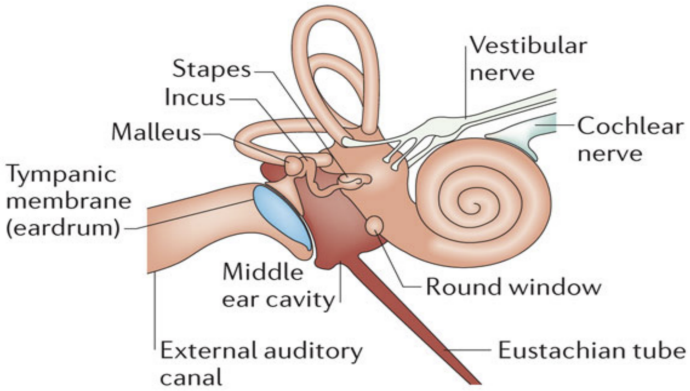

Figure 3: Middle ear and Inner ear pictured together

2.2.1 Fish do not have a tympanic middle ear

Air and water have very different impedances. Most (99%) of air-borne sound energy is reflected at the air-water interface. The middle ear provides impedance matching between air and the fluid-filled inner ear. It does so by collecting sound by

a large tympanic membrane and transmitting it to a smaller-diameter oval window. Fish do not need these adaptations because there is almost no impedance mismatch between water and the inner ear fluid.

In mammals, the three bones of the middle ear (malleus, stapes and incus) form a lever action. As a result, air-borne sounds get transmitted much better to the inner ear. Fish do not need these adaptations because there is almost no impedance mismatch between water and the inner ear fluid.

2.3 Inner ear

2.3.1 Hair Cells are receptor cells of the inner ear

Hair cells are used:

- In the cochlea for hearing
- In the semicircular candle, to sense rotation
- In the sacculus, to sense gravity

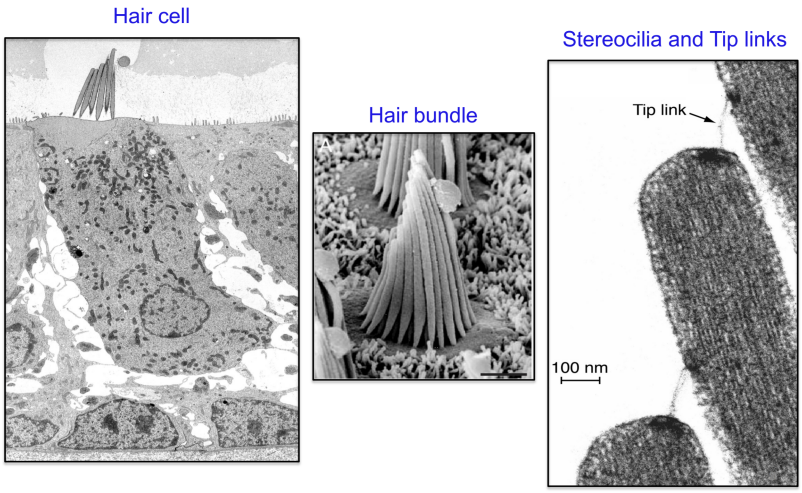

Figure 4: Hair Cell Images

A hair cell is an epithelial cell with a special structure at the top surface. This special structure is a hair bundle, which is composed of many cylinders (stereocilia - basically gigantic mircovilli (microvilli are found on many epithilial cells)) bunched together.

The bottom of each stereocilium is tapered. When the stereocilia are pushed down, they pivot at the bottom and the tip links, as shown in Figure (4) are stretched.

2.3.2 Hair cells in the cochlea

The basilar membrane runs along the spiral organ of Corti in the Cochlea, as shown in Figure 5. The auditory hair cells are located on the thin basilar membrane. These hair cells are responsible for the mechano-transduction of the electrical signal to the brain. These hair cells are pushed up as the basilar membrane oscillates, and this cases them to actively amplify the stimulus, contracting at appropriate times.

The floppy apex of the Basilar Membrane detects high frequencies, while the tense base detects low frequencies. The membrane also has a different thickness and mass along it's length.

2.3.3 Inner ear - active

The inner ear is filled with liquid, which means viscous friction is bound to prohibit passive resonance. The ear is not a passive detector of sounds - it is like an active microphone. Regenerative amplification, the active process, is necessary.

Because the ear is active, it actually produces sounds on its own. These are called Spontaneous Otoacoustic Emissions (SOAEs). The peak SOAE frequencies vary from person to person.

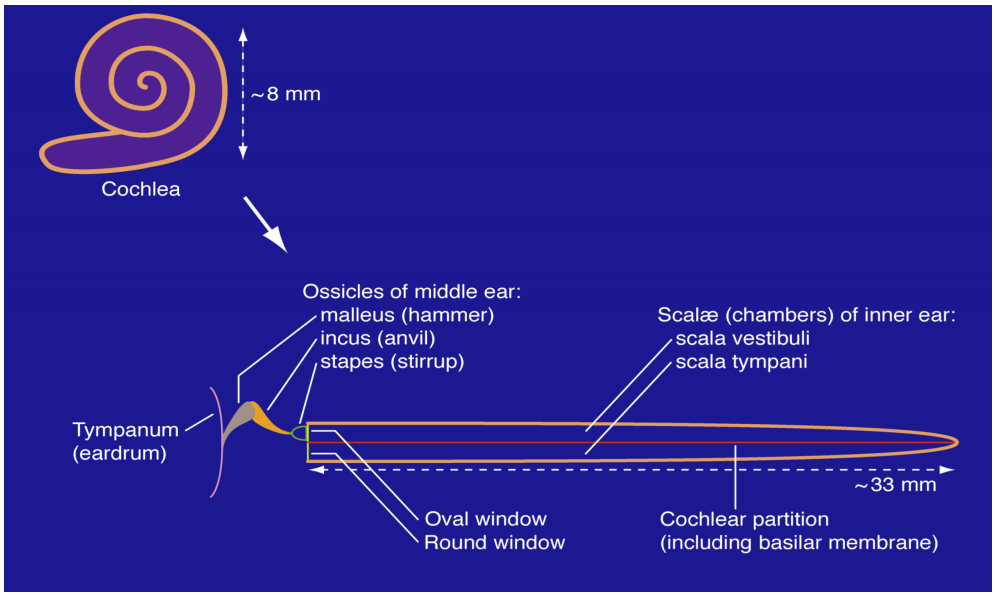

Figure 5: Basilar Membrane in Cochlea

You can also measure the amplitude of a travelling wave in a live, active cochlea vs a dead, passive cochlea and see the amplitude decrease from about 2m to 0.5nm in a Chinchilla specimen.

How is this nm displacement actually measured? This is done using the reflection of a small bead on the cochlea from a laser. The phase shift of the reflection is measured and this is used to calculate the displacement.

2.4 Four characteristics of the ear's active process

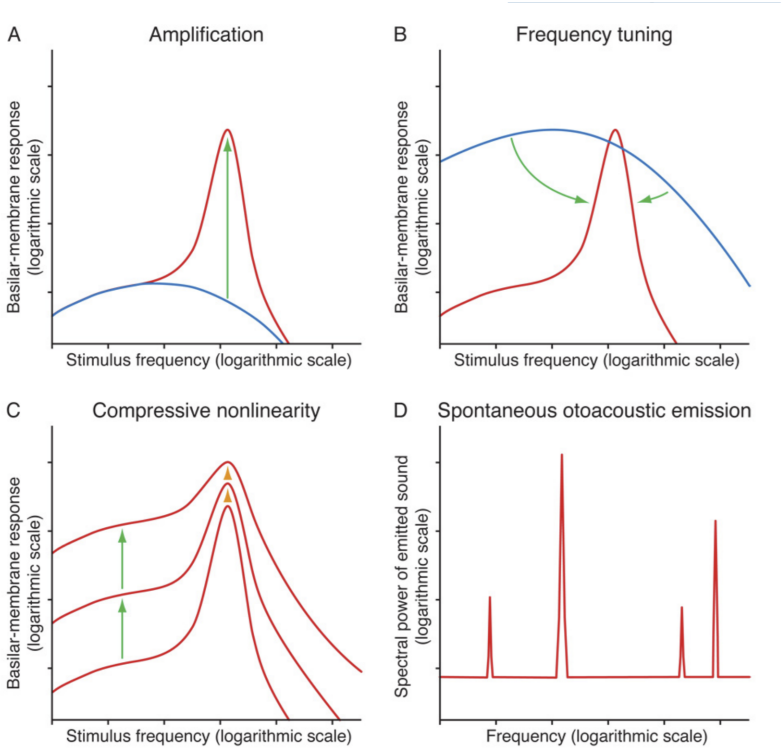

Figure 6: Four characteristics of the ear's active process

1. Amplification
2. Frequency tuning: If you look at a particular point on the basilar membrane, there is a preferred frequency
3. Compressive non-linearity: The drop for a 10 times quieter sound is smaller at the peak. This means that as you decrease the volume of the sound, the active tuning and amplification increases.

3 Mechano-electrical transduction

3.1 Classical gating spring model

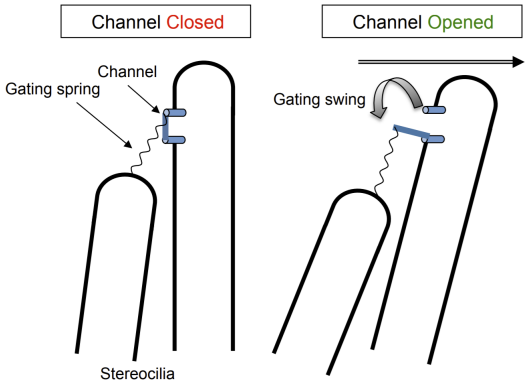

Figure 7: Illustration of Classical gating spring model

By this model, the probability of the spring being open is a sigmoid of the relative distance between the current and resting positions of the stereocilia.

$$p(\text{open}) = \frac{1}{1 + e^{\frac{-z(X - X_0)}{kT}}} \tag{1}$$

z is the **single channel gating force** and is given by $k_{\text{gs}}d$ where k_{gs} is the gating spring stiffness and d is the gating swing. z is approximately several pN and the displacement is around several nm. The energy is a few kT.

In addition, it may be useful to remember that k is the Botzmann constant and so kT is the thermal energy unit.

3.1.1 How is energy injected into the system?

Ca^{2+} - dependent channel re-closure exerts mechanical force.

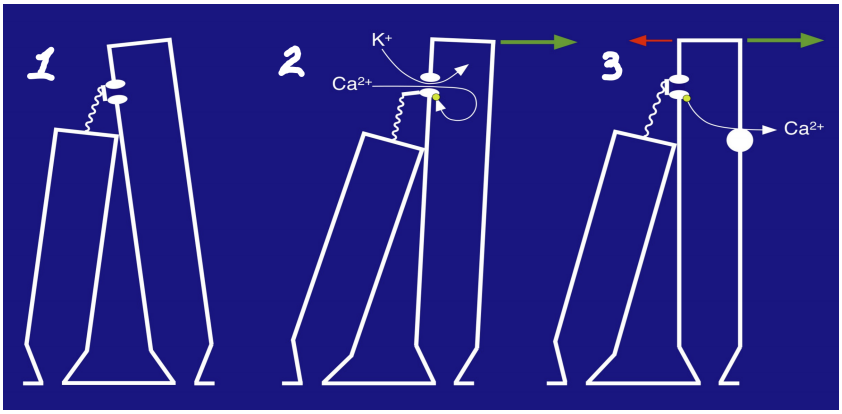

Figure 8: Illustration of how Ca^{2+} - dependent channel re-closure exerts mechanical force.

1. The hair bundle is at it's resting position
2. A sinusoidal sound wave stimulus (shown in green) comes along and pushes the stereocilia, opening the channel. The majority of the current into the channel is carried by potassium. This is because Stereocilia are surrounded by endolymph which is rich in potassium and low in sodium. The channel closes once calcium binds to the channel or close to the channel.

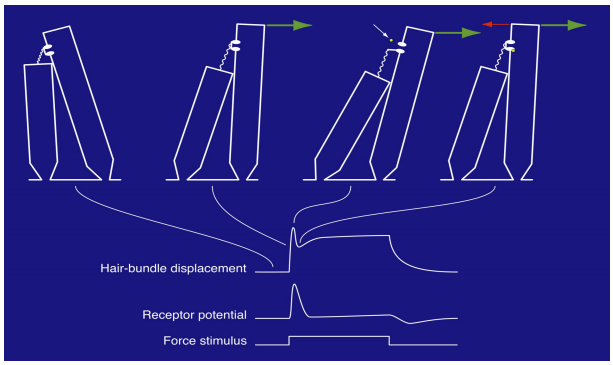

Figure 9: Illustration of Fast Adaptation

3. The spring is under more tension and this produces a backwards mechanical force (shown in red).

The mechanical force (shown in red) produces "Fast Adaptation". This can be tested experimentally.

One may ask the question of how much work does the binding and unbinding of calcium actually do? When the channel is closed, the resting concentration of calcium is small. This is maintained by calcium pumps / buffers. $[Ca^{2+}] = 0.05\mu M$ throughout the tip.

When the channel opens, $[Ca^{2+}] = 37\mu M$ within 5nm of the channel.

For one cycle, the change in Gibb's Free Energy, ΔG, is given by $kT \ln(\frac{[Ca^{2+}]_2}{[Ca^{2+}]_1}) = kT \ln(\frac{37}{0.05}) \approx 30zJ \approx 7kT$

7kT has a measurable effect on the mechanics of the hair bundle. Fast adaptation can power the active process on a cycle by cycle bases and can power the active process even at high frequencies. The calcium binding aligns with motion and specifically tuned to the primary frequency in order to amplify that frequency.

3.1.2 How does the mechanotransduction apparatus minimize viscous energy dissipation?

The hair bundle is small, soft, light and immersed in viscous endolymph. The Reynolds number (ratio of intertia to viscous forces) is around 10^{-4}, which means very viscous. However, the hair bundle minimises viscous friction by not letting the liquid between the stereocilia move. If the liquid doesn't move, there is no viscous friction.

The extent of liquid movement between stereocilia depends on how much stereocilia dynamically separate from one another. Measuring the relative motions between stereocilia requires a sensitive technique. This is because the gap between stereocilia is around 50nm while the wavelength of visible light varies from around 400-750nm. This means you cannot measure the movement using a camera. Instead, another optical tool is used, called a **dual-beam laser interferometer**.

The dual-beam differential interferometer uses two coherent light sources to monitor hair-bundle movements at two positions simultaneously. The dual-beam laser interferometer, measures hair bundle movements with subnanometer spatial resolution and submillisecond temporal resolution.

With interferometers, the inference pattern is a function of the phase lag between 2 beams, 1 of which is passing through the moving object. If you measure the movement at multiple places with multiple interferometers with different wavelengths of light, you get a dual-beam interferometer.

The hair bundle moves as a unit at a scale of several dozen nanometers. The relative mode is much smaller than the common mode (this is shown by the high coherence between the red and green beams in Figure (10)). But, to understand the principles governing energy dissipation, one has to quantify the **relative** mode.

The hair bundle's mechanical nonlinearity can be exploited to measure the tiny, but highly dissipative relative mode. Hair-bundle stiffness is nonlinear because of channel gating. Around its resting position, the hair bundle stiffness is highly non-linear. When the hair bundle is stimulated mechanically at two frequencies f_1 and f_2, it will produce distortion products at $2f_1$, $f_1 + f_2$, $2f_2$, etc. The distortion product amplitude depends on the internal degrees of freedom.

Figure (11) shows the results of applying frequencies f_1 and f_2 to a hair bundle using a glass pipette and then measuring the relative motion between the short and long edges using dual beam interferometry. Quadratic distortion products at the short edge are $\approx$ 5nm relative to the tall edge. This means there is around 5nm relative motion across the bundle, which is about 7 stereocilia gaps. The upper limit on movement between adjacent stereocilia is 0.7nm, the size of 1 water molecule - very small. This means there is **no flux of water in and out of an array of stereocilia** which means no viscous dissipation in the hair bundle.

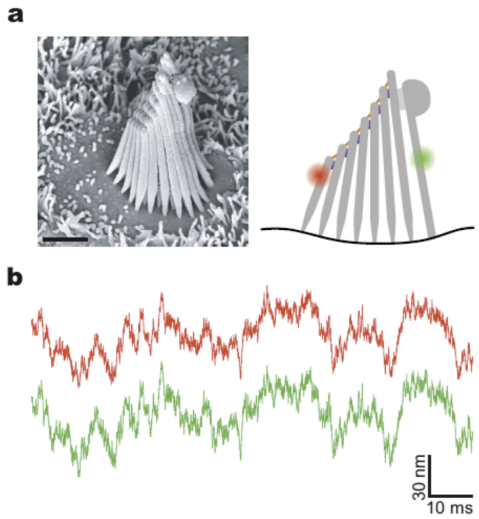

Figure 10: Illustration of dual-beam laser interferometer

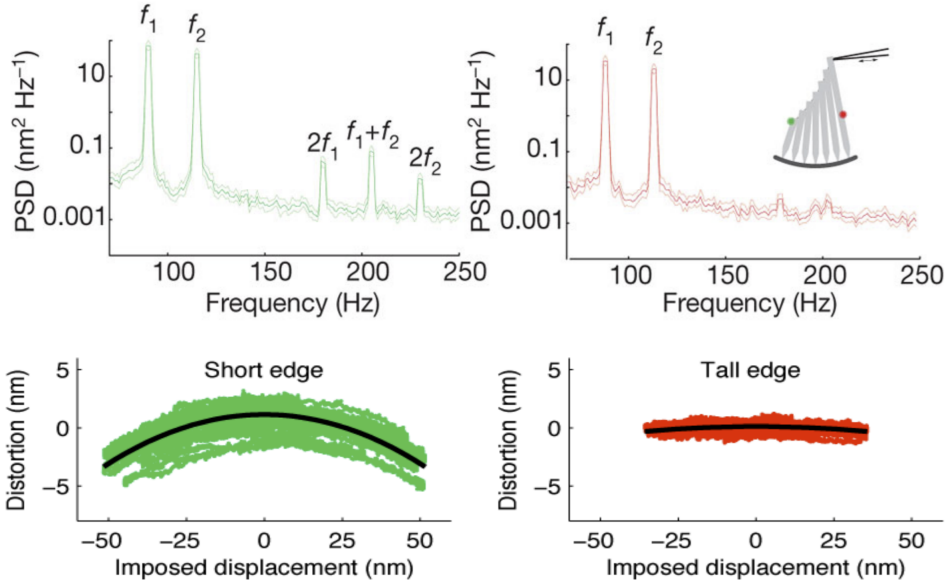

Figure 11: Relative motions at distortion frequencies are less than 5 nm between the hair bundle edges.

Finite-element modelling of the hydrodynamic and elastic couplng between stereocilia has shown that hydrodynamic coupling increases with frequency, which means that the stereocilia all move together at high frequencies (even without tip links). This means that viscous drag decreases with frequency.

In fact, tip links fail to make all stereocilia move together. Instead, they cause additional relative modes within the array, decreasing coherence. However, besides tip links there are also **horizontal top connectors**. Horizontal top connectors efficiently couple all stereocilia, even at low frequencies. All of this means that dissipation occurs mainly in the outer boundary layer and inner stereocilia contribute negligibly to the total drag. In fact 60 stereocilia moving together incur the same amount of viscous drag as 3 moving individually.

3.2 Difference between slow and fast adaptation

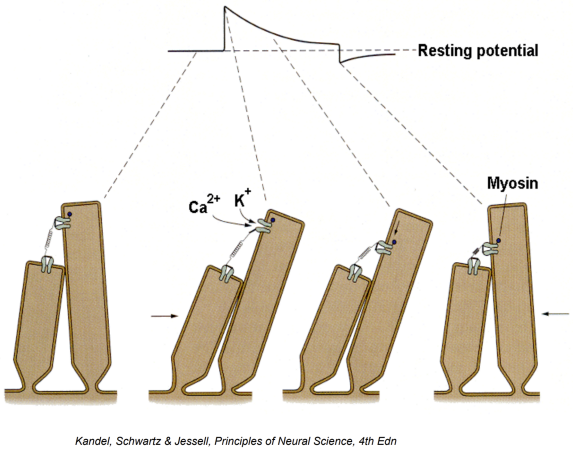

Figure 12: Slow adaptation

The force exerted by myosin molecules is the basis of slow adaptation. As shown in the third stage in Figure (12), as the channel closes it also moves down because myosin motors let go after Ca^{2+} enters.

3.3 Questions about classical model

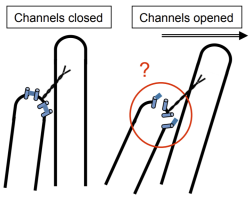

Figure 13: Modern gating spring model

- Classical model: One channel at the upper end of the tip link
- Modern model: Two channels at the lower end of the tip link

The gating swing is variable and very large. This raises a couple of questions. Firstly, how can the same channel swing so differently? Secondly, how can a 5nm channel perform a 25nm swing?

Ion channels embedded in the lipid bilayer introduce curvature, which causes hydrophobic mismatch. When a channel is closed, the hydrophobic mismatch is at its lowest, while when it opens, the mismatch is at its highest. This means that closed and open channels experience a repulsion, while two open channels experience the greatest attratcion. The length scale of this interaction is several nm. The cycle of both channels open to both channels closed and back produces cyclical lateral motion of the channels, modulating tension on the tip link.

In summary, gating swing seems too great to originate within a MET channel. It's more likely a collective phenomenon (2 MET channels + lipid bilayer).

4 Sound localization

- Interaural time difference (ITD) is used for sound localization in the horizontal plane
- Phase locking in the auditory nerve (afferent fibers)
  - Coincidence $\longrightarrow$ phase locked inputs $\longrightarrow$ output spikes of coincidence detector neuron
  - No coincidence $\longrightarrow$ no output spikes
- Jeffress model of coincidence detection for sound localization

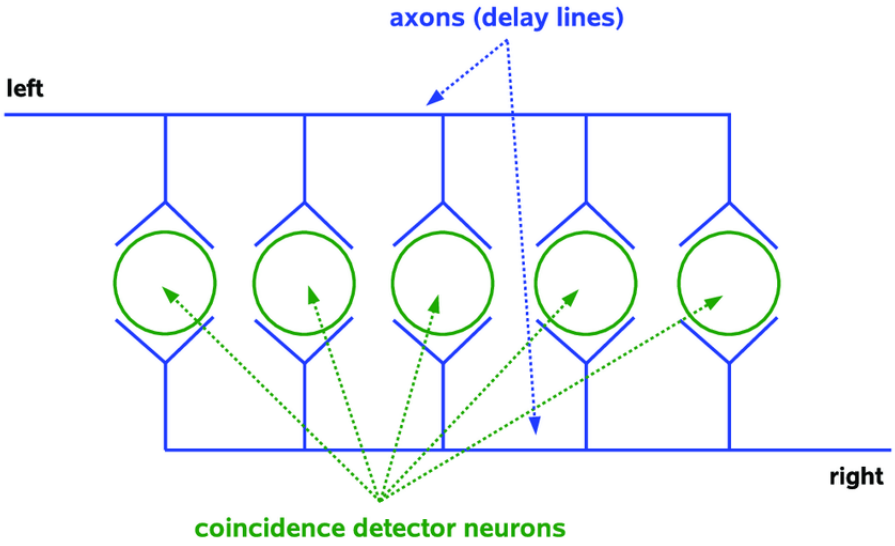

Figure 14: Jeffress model diagram

- Calculation of ITD requires a coincidence detection of inputs from the right and left cochlear nuclei by Medial Superior Olive (Or Lam in birds) neurons.
- **Single tones cannot be localized unambiguously using ITDs.** This is due to phase ambiguity - when you mix different frequencies, there is a clear peak for broad band sounds, but multiple peaks for a single tone.
- Not only is Interaural time difference used, but also the intensity difference is used as a 2nd binaural cue. This is called interaural level difference (ILD). This is a simple mechanism: no delay lines, no phase-locking, just a balance between excitation and inhibition.
- ILD involves a competition between a direct excitation and an indirect (di-synaptic) inhibition from the two cochlear nuclei.
- A common theme in sensory neuroscience is the existence of parallel neural pathways, each analysing a specific aspect of a complex sensory stimulus. For example, in the visual cortex, different neuronal pathways analyse colour and motion. In the auditory system, the Medial superior olive processing ITDs and lateral superior olive processing ILDs is an example of parallel pathways used specifically for sound localisation.
- Auditory space-processing pathway in mammals/birds, important shared principles:
  - Parallel pathways for ITDs and ILDs that converge in the Inferior Colliculus
  - Tonotopy is preserved in the pathway to the coretx but it is discarded in the pathway to SC/OT
  - Auditory space map in Superior Colliculus/Optic Tectum

5 Cortical processing

The auditory cortex (AC) displays the properties of selectivity and invariance, which are properties of the central sensory ssytems.

- The AC is tonotopically organised. This can be shown by testing which neurons prefer which frequencies electrophysically or optically. Tonotopy, established in the cochlea, is **preserved** in the thalamus and A1.

- Tuning curve of each neuron is v-shaped - each neuron has a preferred frequency.
- The AC is plastic. Especially in young animals. This comes from the fact that in addition to auditory inputs from the periphery, many areas also project the cortex and modulate the response of auditory neurons. For example Nucleus basalis sends neuromodulator (Acetylcholine) to many areas including AC. Released when animal engaged in an activity. This can be used to re-tune neurons to respond to other frequencies.
- Selectivity: weak inputs together (AND-like operation) activate the cell
- Invariance: strong inputs individually (OR-like operation) activate the cell
- Biological neurons can perform possible more than one operation, while in computer neural networks, a neuron performs one fixed operation
- How to determine what operation a neuron performs? Can use "Summation Index"

$$SmI = \frac{R_{A+B} - \text{MAX}(R_A, R_B)}{\text{MIN}(R_A, R_B)} \tag{2}$$

  - SmI - max rule
  - not 0 - tuning
  - > 1 - supralinear additivity
- Distribution of Summation Index is broad, indicating a wide range of operations, and has a peak at zero that corresponds to the MAX function. SmI distributions are similar in the auditory and visual systems: both are broad and peak at zero.
- Auditory neurons display the same feature recombination functions as in the standard model of primate ventral stream, notably the max response
- Adaptation acts as a swicth for these functions OR $\longrightarrow$ AND
- Biophysically, this finding is consistent with the theoretical flexibility of the canonical microcircuit, and is expected because of threshold nonlinearity
- However, this is incompatitble with any model that relies on a fixed mapping between neurons and computations (aka all common modern techniques)
- Auditory cortical neurons represent statistical regularities in natural stimuli

6 Vocal learning in songbirds and in children

- Vocal learning is the best studied example of auditory plasticity. It is not the same as auditory learning, which is more common.
- Vocal learners have specialised neural circuits which are homologous in birds and mammals.
- Vocal production learning involves three systems:
  1. Sensory (representations of natural stimuli) - generic system
  2. Motor (representations of learned vocal gestures) - special in Vocal Learners
  3. A feedback system to correct errors - special in Vocal Learners (Cortico-striatal-thalamic loop)
- Auditory cortical neurons represent statistical regularities in natural stimuli (e.g., birdsong) cortex is.
- If you compare biological and artificial NNs, you can see that biological NNs also perform statistical optimisation
- Vocal learning evolved in parallel between birds and mammals
- Phases of vocal learning
  1. Sensory
  2. Sensorimotor

6.1 Auditory illusions; seeing with the auditory cortex in the blind.

Blind people can use sounds to see the world (using visual cortex). This illustrates how plastic cortex is.

7 Auditory specialists: bats and whales

- Bats have the most sensitive mammalian ears - they can hear sounds as quite as -20dB, where 0dB is the quietest a human can hear
- Bats can also echolocate
- 2 classes of echolocating bats:
  - FM bats (frequency modulated call - chirp high to low - good to measure targets in open ear where the only small objects are insects). Bats listen to time difference between call and echo,
  - CF-FM bats - pulse made up of constant frequency harmonics followed by an FM chirp. The CF pulse is used to measure the target velocity using the doppler shift, This is useful to detect flying insects an a dense forest.
- Bats have a specialised structure in the brain. The DSCF area ("auditory fovea") contains neurons with extremely narrow frequency tuning, centered around the dominant harmonic of the bat call (CF_2, the 2nd harmoic). This is the narrowest frequency tuning in any animal's cortex.
- Hyper-selectivity in bat auditory cortex - there is an over representation of the second harmonic echo in the DSCF area in the bat's auditory cortex. When a bat tries to get the velocity of a moving insect, it will adjust its call until to Doppler shifted echo is of a frequency of CF_2 so that they can use their "auditory fovea" effectively.

8 Hearing in insects; fundamental principles shared by insect and vertebrate hearing

Different groups of insects developed hearing in parallel, much before mammalian hearing. Insect ears look similar to mammalian ears on a microscopic level due to the channel spring attachments, however on a macroscopic level they are completely different pieces of hardware. For example, insect ears do not have hair cells.

Insect hearing is also active. Whenever anything oscillates spontaneously, this is evidence of the active process.

The same mathematical model of an active nonlinear oscillator can be applied to both vertebrates and insects. This model is the Hopf bifurcation, represented by the equation:

$$\frac{dz}{dt} = \mu z + i\omega_0 z - |z|^2 z \tag{3}$$

Where z is a complex variable whose real part represents displacement. The effect of the individual terms is illustrated in Figure (15).

The control parameter in the equation is μ. The control parameter is the key for any bifurcation, which is when a system suddenly changes upon a change in the control parameter, as shown in Figure (16).

At the Hopf bifurcation, μ, the control parameter, assumes a critical value, and the quiescent, non-oscillatory state becomes unstable. Above this critical value, the oscillator displays self-sustained oscillations whose amplitudes and frequencies depend on μ.

At the characteristic frequency, sensitivity nonlinearly increases, in a log-log plot, with a slope of -2/3. At frequencies far away from this characteristic frequency, however, sensitivity stays largely constant across intensities. The result is a frequency-dependent amplification, which nonlinearly boosts the sensitivity to sounds at, or near, the characteristic frequency.

The softer the sound (i.e. 0 dB), the sharper becomes the frequency-tuning, resulting in an increased frequency selectivity.

The ears of the human and of the fruit fly (Drosophila melanogaster) are macroscopically very different. Yet from the dynamical systems perspective, the four characteristics of the active process in both species can be elegantly described by the same principle. Describing the ear as an active nonlinear oscillator, operating at the Hopf bifurcation, can explain the four characteristics of the active process both in vertebrate ears and in insect ears. Far from the bifurcation, when the control parameter is large and negative, the system exhibits linear behaviour. Close to the bifurcation, when the control parameter is close to zero, the system displays amplification and compressive nonlinearity (greater amplification of smaller stimuli) when stimulated at the natural frequency ω_0. Past the bifurcation, when the control parameter is positive, the system displays limit cycle oscillations, which may underlie spontaneous otoacoustic emissions.

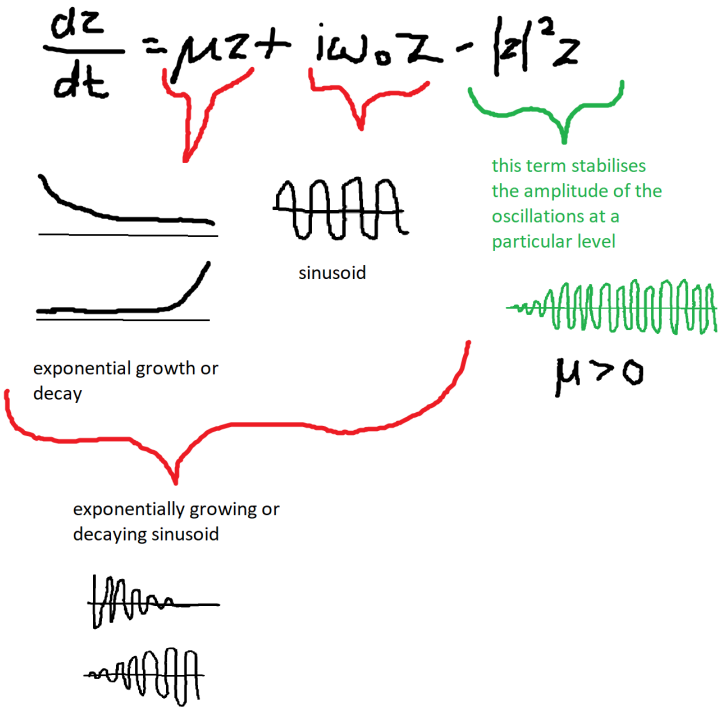

Figure 15: Diagram illustrating the terms of the characteristic equation of the Hopf bifurcation

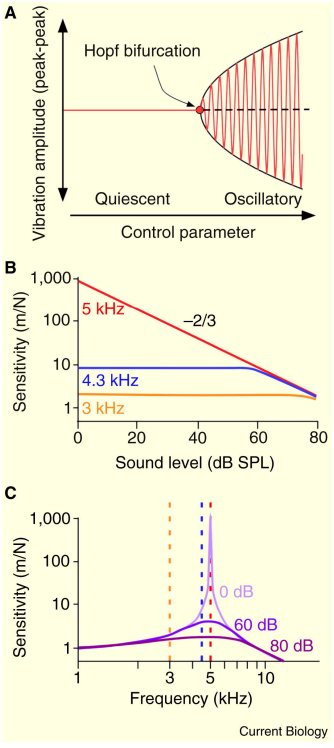

Figure 16: The critical Hopf oscillator

9 Fundamentals of audio processing

9.1 Speech as a complex waveform in time

- A speech signal can be represented as a temporal waveform or spectrogram
- The spectrogram has Time on the x axis, Frequency on the y axis, an Intensity represented using a colour code
- Speech can be analysed using the Fourier Transform, which breaks the signal down into it's constituent sinusoids
- For speech signals to be processed digitally, they must be converted from from continuous to discrete, which is a process called "sampling"
- Nyquist criterion: A continuous signal $x(t)$ with frequencies no higher than or equal to $f_{\max}$ can be reconstructed exactly from its samples $x_n = x(nT_s)$ if the samples are taken at a rate $f_s = 1/T_s$ from which $f_s \geq 2f_{\max}$
- A violation of the sampling theorem results in folding of signals at higher frequencies into signals at lower frequencies, corrupting the sampled signal. This is called **alisaing**.
- To avoid aliasing, the signal has to be low pass filtered before sampling.
- The **Discrete Fourier Transform** is a Fourier transform of signals which are sampled and time-limited. If the signal is time- and band-limited it can be represented by a discrete spectrum (amplitude and phase) across a finite number of frequencies which are multiples of the fundamental frequency (the inverse of the signal duration).
- The **Short-time Fourier Transform (STFT)** is useful seeing as a speech signal has many temporal fluctuations and we often want to know the spectrum within a certain time window. To do this we multiply the speech signal with a windowing function before performing the DFT. However, there is a trade off seeing as a small window width provides high temporal resolution in exchange for low frequency resolution. This is because the finite frequency resolution Δf is approximately the inverse of the temporal resolution Δt.
- A spectrogram uses many such time windows, centered at different time points, to analyze the signal.

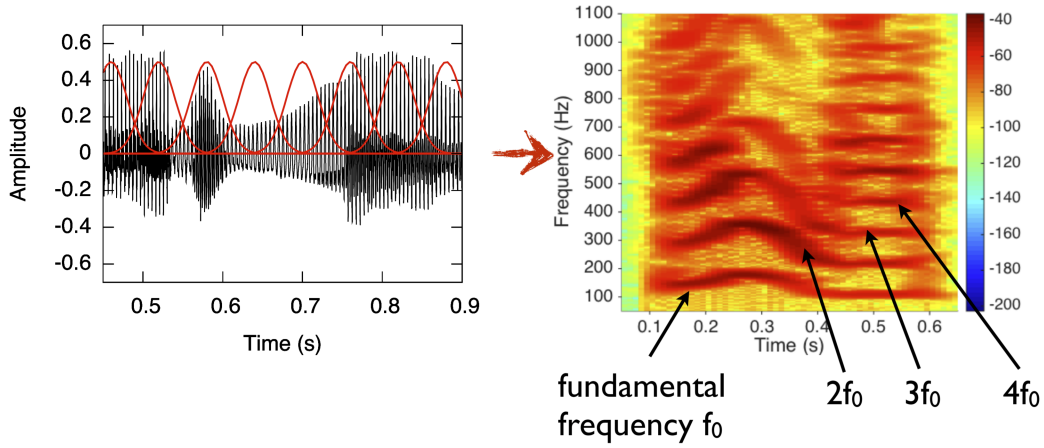

Figure 17: Here a spectrogram reveals the fundamental frequency of speech

- The ideal windowing function gives a good frequency resolution and surpression of sidelobes.

10 Speech production

- The pitch of the voiced parts of speech arises from the vibration of the vocal folds
- Vibration of the vocal folds can be seen from the vocal tract using videostroboscopy
- The vibration of the vocal folds produces voiced speech with a fundamental frequency and higher harmonics
- The vocal tract and nasal cavities introduce characteristic resonances (formants). Modification of these cavities changes the formants; we perceive the resulting sound as distinct vowels.
- Formants are resonances in the spectrum as shaped by the vocal tract. They are labeled by their frequency ranking (from low to high), e.g. F_1, F_2, F_3 etc.

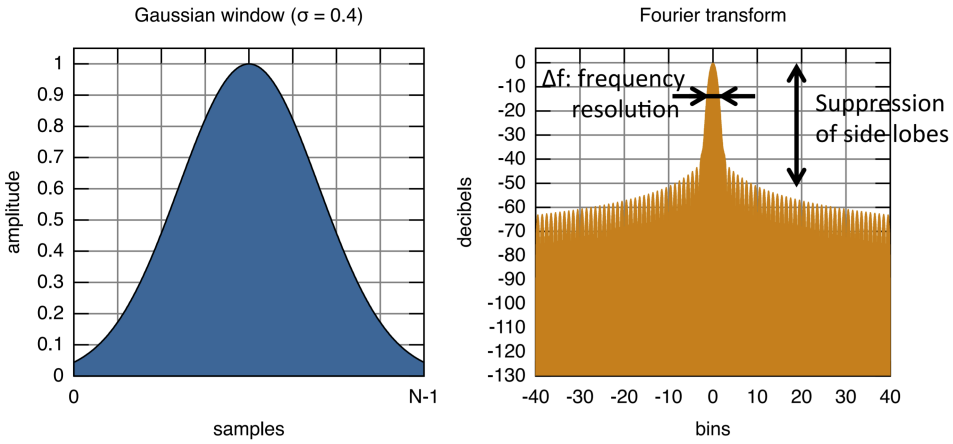

Figure 18: Fourier transform of a windowing function

- Some parts of speech are voiceless, such as some consonants
- A condition called **Aphasia** means that one has trouble with speech production. Much about the neural substrate for speech processing has been learned from patients with lesions and their resulting difficulties with language.
- **Broca's area** in the brain plays a major role in producing speech
- **Vernicker's area** is used for comprehension

11 Linear prediction and speech coding

A tube model for the vocal tract can be used to describe speech production:

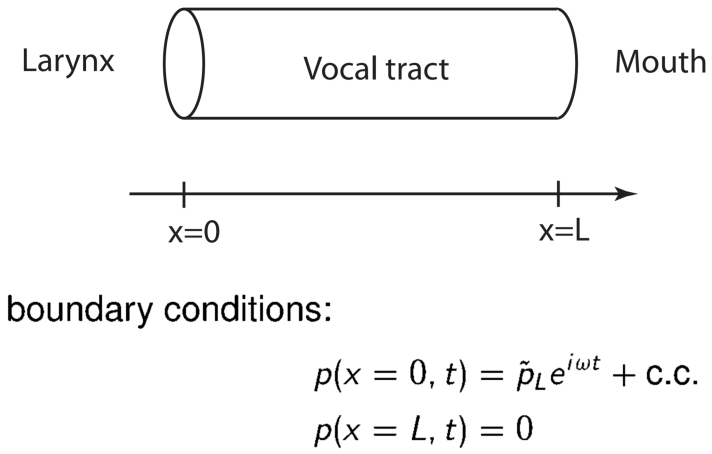

Figure 19: Tube model for the vocal tract

The equation for sound propagation is:

$$\rho_0 \kappa \frac{\partial^2 p}{\partial t^2} = \frac{\partial^2 p}{\partial x^2} \tag{4}$$

This can be solved using the Ansatz for travelling waves:

$$p = \tilde{p}_f e^{i\omega t - ikx} + \tilde{p}_b e^{i\omega t + ikx} + \text{ c.c.} \tag{5}$$

Where $\tilde{p}_f$ is the amplitude of the forward propagating wave and $\tilde{p}_b$ is the amplitude of the backward propagating wave (c.c. means complex conjugate and is needed to make pressure real).

- ω: angular frequency

- T: period
- f: frequency
- k: wavenumber
- λ: wavelength
- κ: compressibility
- ρ: density

Important results:

- Dispersion relation: $k = \sqrt{\rho_0 \kappa}\omega$ $(k > 0)$ - this is the relation between the spacial and temporal period
- Sound velocity $c = \frac{\omega}{k} = \frac{1}{\sqrt{\rho_0 \kappa}}$
- $\tilde{p}_b = \tilde{p}_L - \tilde{p}_f$

$$\tilde{p}_f = \frac{\tilde{p}_L}{1 - e^{-2ikL}} = \frac{\tilde{p}_L}{1 - e^{-2i\omega L\sqrt{\rho_0 \kappa}}} \tag{6}$$

  Which leads to resonance when the denominator is equal to 0, namely when $\omega = \frac{n\pi}{L\sqrt{\rho_0 \kappa}}$

Using the **z-transform** $(p_f(z) = \sum_{\infty}^{\infty} p_f^{(m)} z^{-m}$, we can express Equation (6) as:

$$p_f(z) = \frac{p_L(z)}{1 - z^{-2LF_s\sqrt{\rho_0 \kappa}}} \tag{7}$$

where F_s is the sampling frequency.

However, the real vocal tract resembled several connected tubes, which means the equation should contain multiple (n) resonances:

$$p_f(z) = \frac{p_L(z)}{1 - \sum_l a_l z^{-l}} \tag{8}$$

From this it follows that we can use linear predictive coding to predict the m^{th} sample from previous samples in a linear manner, using the equation:

$$p_f^{(m)} = p_L + \sum_l p_f^{(m-l)} \tag{9}$$

In essence, linear predictive coding encodes the speech signal through the coefficients a_l that describe the resonances of the vocal tract. This can be a more efficient format in which to store speech in a computer.

12 Pitch determination

Hundreds of methods for pitch determination exist, none of which are error-free. The problem is difficult because:

- Speech is nonstationary. Pitch and the temporal waveform change constantly.
- Some voiced segments last only a few pitch periods.
- The fundamental frequency can vary over a wide range, from 50 to 800 Hz.
- In telecommunication systems the signal is often band-pass filtered, say between 300 - 3000 Hz, and the fundamental frequency itself may be removed from the signal.

Broadly, methods of pitch determination can be split into time domain and frequency domain methods.

12.1 Autocorrelation

Autocorrelation is a time domain method. Given a speech signal $s(t)$ with a mean of 0, and a pitch period of T_0, the autocorrelation is given by:

$$r(\tau) = \frac{1}{T}\int_0^T s(t)s(t+\tau)dt \tag{10}$$

The autocorrelation has a maximum at the delay $\tau = 0$ and another maximum at $\tau = T_0$.

For a discrete speech signal with sampling rate F_s, we used the autocorrelation for a discrete signal, which is given by:

$$r(\tau) = \frac{1}{N}\sum_{i=1}^{N} s_i s_{i+F_S\tau} \tag{11}$$

12.2 Average magnitude distance function

AMDF is also a time domain method.

$$\text{AMDF}(\tau) = \frac{1}{N}\sum_{i=1}^{N} |s_i - s_{i+F_S\tau}| \tag{12}$$

The AMDF has a minimum at delay $\tau = 0$ and another minimum at $\tau = T_0$.

12.3 Hilbert-Huang transform

The pitch of a speech signal can also be determined from the Hilbert-Huang transform. To this end, the Hilbert-Huang transform decomposes the speech signal into different intrinsic mode functions. The number of extrema and the number of zero crossings of the instinsic mode functions are either equal or differ by one.

12.4 Problems with time domain methods

Methods in the time domain yield problems when:

- A strong first formant appears at the second or third harmonic (results in a false prediction of period as $T_0/2$ or $T_0/2$)
- The fundamental frequency is missing due to band-pass filtering (results in a false prediction of a higher period, such as $T_0/2$ or $T_0/2$)

12.5 Cepstrum method

The power cepstrum is given by

$$\text{power cepstrum}(\tau) = |\text{F}^{-1}\{\ln|\text{F}[s(t)]|\}| \tag{13}$$

τ can be called the lag, pseudo-time, or "quefrency".

The power cepstrum has maximum at lag $\tau = T_0$.

12.6 Problems with frequency domain methods

Methods in the frequency domain are not resilient to white noise, unlike time domain models.

13 Speech recognition with hidden Markov models

13.1 Phones and phonemes

Definitions:

- Phonology: Branch of linguistics that investigates the organization of sound in a language
- Phoneme: Basic unit of a language's phonology; the smallest contrastive linguistic unit which may bring about a change of meaning in a certain language (language dependent)
- Phone: Smallest speech segment that possesses distinct physical or perceptual properties, independent of its role in language (language-independent, mostly used for speech recognition)

Multiple phones that correspond to the same phoneme in a language are called "allophones".

13.2 Mel frequency scale

The mel scale is a scale of pitches judged by listeners to be equal in distance one from another. This means it is a perceptual scale. The reference point between this scale and normal frequency measurement is defined by equating a 1000 Hz tone, 40 dB above the listener's threshold, with a pitch of 1000 mels.

Frequencies that are equidistant on the mel scale are perceived to differ by the same amount in their pitch.

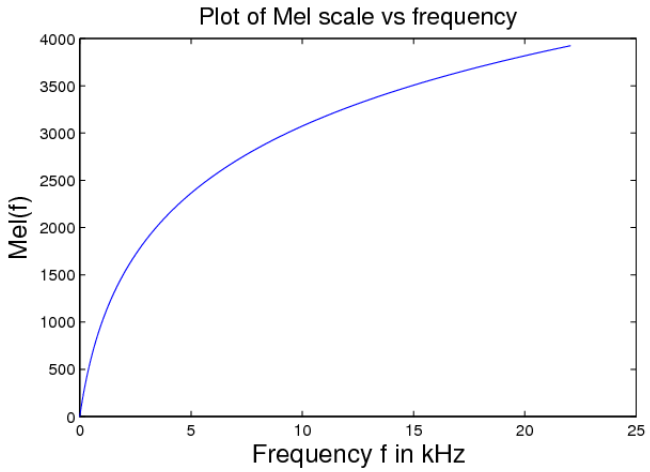

Figure 20: Mel frequency scale

13.3 Mel frequency cepstral coefficients (MFCCs)

13.3.1 Normal Cepstral coefficients

As in Equation (13), the power cepstrum is given by $|\mathrm{F}^{-1}\{\ln|\mathrm{F}[s(t)]|\}|$. Therefore to get the cepstral coefficients, we:

1. take Fourier transform
2. take log of the amplitude
3. take inverse Fourier transform
4. take amplitude

13.3.2 Mel-frequency cepstral coefficients

1. take Fourier transform
2. **map amplitudes on a mel scale (using windowing)**

3. take log of the amplitude
4. take inverse Fourier transform
5. take amplitude

13.4 Speech recognition introduction

- We divide speech into a set of phones q_1, q_2, q_3... and then attempt to recognize the individual phones
- Each phone is characterized by its Mel-frequency cepstral coefficients (MFCCs) $\mathbf{y}^{(q)} = (y_1^{(q)}, y_2^{(q)}, y_3^{(q)} \ ... \ y_n^{(q)})$
- The task is to identify the phone by its MFCCs at a certain time. The idea here is to use a maximal-likelihood approach - we want to find the phone q that that maximizes the probability that q corresponds to the observation $\mathbf{y}$. This means maximising $P(q|\mathbf{y})$.
- However, we may be able to compute $P(\mathbf{y}|q)$ but not $P(q|\mathbf{y})$ directly. In other words, we can compute the probability of observing those particular MFCCs given a certain phone occured, but cannot directly computer the probability of a certain phone given the MFCCs.
- Here we can apply Bayes Theorem, which states that

$$P(q|\mathbf{y}) = \frac{P(\mathbf{y}|q)P(q)}{P(\mathbf{y})} \tag{14}$$

- Therefore, the problem can be reframed as maximising $P(\mathbf{y}|q)P(q)$. Now all we need is to compute $P(\mathbf{y}|q)$ which can be done using a **Hidden Markov Model**.

13.5 Hidden Markov Model

13.5.1 What is a Markov Process

A Markov process is a random process in which the future is independent of the past, given the present. Thus, Markov processes are the natural stochastic analogs of the deterministic processes described by differential and difference equations. They form one of the most important classes of random processes.

$\mathbf{x}$ is a sequence of random variables $\mathbf{x} = (x_1, x_2, ..., x_n)$ where each x_i is one of the states 1 to M and the probability that state x_i transitions to state x_j is a_{ij}.

13.5.2 Hidden state

Each state k produces output y_j with probability $b_k(y_j)$. **Only the outputs can be observed**, which means the states x_i are "hidden".

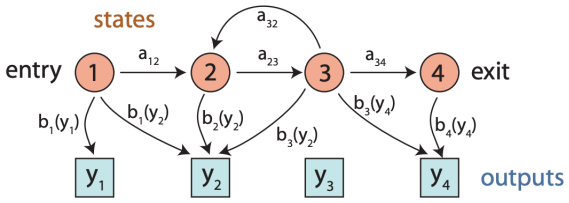

Figure 21: Hidden Markov Model - an example diagram

13.5.3 Three fundamental problems

A HMM is described with the parameters $\theta = (\vec{\pi}, A, B)$

- $\vec{\pi}$ is a vector of length M, where M is the number of states, and represents the starting probability of each state.
- A is a $M \times M$ matrix whose entries a_{ij} represent the transition probability between state i and state j.

- B is a $M \times N$ matrix, where N is the number of observable outputs, whose entries $b_i(y_i)$ represent the observation probability of output j from state i.

1. **Likelihood:** Given an HMM with parameters θ and an observation sequence of states observed $\vec{y}$, determine the likelihood of obtaining the observations (determine $P(\vec{y}|\theta)$.
2. **Decoding:** Given an HMM with parameters θ and an observation sequence of states observed $\vec{y}$, discover the sequence of hidden states $\vec{x}$ that is most likely to have produced the observations.
3. **Learning:** Given an observation sequence $\vec{y}$ and the set of states in the HMM, learn the HMM parameters θ.

13.6 Application of HMM to speech recognition

Each phone q can be described by a distinct HMM.

We want to determine the probability of observing the MFCCs $\mathbf{y}$ at a given time, under the assumption that they result from the utterance of phone q, which means this is problem 1, "likelihood".

$$P(\mathbf{y}|q) = \sum_x P(\mathbf{y}|x)P(x) = \sum_x P(\mathbf{y}, \mathbf{x}) \tag{15}$$

1. The number of sequences of hidden states that can give rise to the observations (and over which the sum is carried out) can be very large. This means we need an efficient algorithm for determining the likelihood. This is called the Forward algorithm.
2. We don't know the HMM for the phone q. This means we need to determine the parameters of the HMM from many observations $\mathbf{y}$ that are known to correspond to the phone q. This is problem 3, "learning".
   - Mathematically, the problem of learning can be described as finding the parameters θ^* that maximise the probability of observing the MFCCs $\mathbf{y}^{(q)}$ of phone q.
   - This can be solved using the Baum-Welch algorithm.

13.6.1 Plan for speech recognition through HMMs

- For each phone q, train a HMM through MFCCs $\mathbf{y}^{(q)}$ that are known to correspond to that phone.
- To match phones to an unknown speech sample:
  1. Divide the speech sample into short segments that correspond to individual phones and extract the MFCCs $\mathbf{y}$
  2. For each phone q, compute the probability $P(\mathbf{y}|q)$ of observing the MFCCs $\mathbf{y}$ under the assumption of phone q (using the HMM for phone q)
  3. Determine for which phone q the product $P(\mathbf{y}|q)P(q)$ is maximal
  4. Associate this phone to the MFCCs $\mathbf{y}$

14 Speech recognition with neural networks

14.1 Perceptron

In machine learning, the perceptron is an algorithm for supervised learning of binary classifiers. A binary classifier is a function which can decide whether or not an input, represented by a vector of numbers, belongs to some specific class. It is a type of linear classifier, i.e. a classification algorithm that makes its predictions based on a linear predictor function combining a set of weights with the feature vector.

The perceptron can be used to classify vowels, for instance, based on formants (or mel-frequency cepstral coefficients).

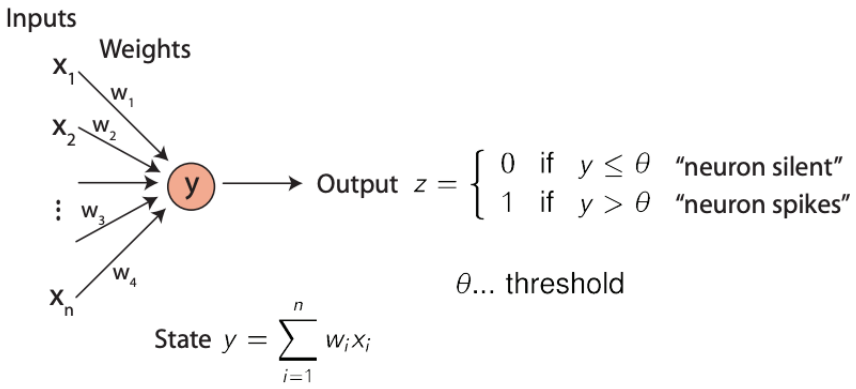

Figure 22: Perceptron

14.1.1 Training the weights of the Perceptron

We train the weights $\mathbf{w}_i$ such that classification is as accurate as possible.

We get training data in the form $\{(\mathbf{x}^{(1)}, d^{(1)}), (\mathbf{x}^{(2)}, d^{(2)}), \ \ldots \ , (\mathbf{x}^{(N)}, d^{(N)})\}$. Each $\mathbf{x}^{(i)}$ is the input from training set i, while $d^{(i)}$ is the output from training set i, namely is it class 0 or 1.

Algorithm for training:

1. Initialize weights at 0 or at a (small) random value, initialize training step k as $k = 0$
2. For each training set i do the following repeatedly:
   (a) Calculate the perceptron's output $z^{(i)}(k)$ from input $\mathbf{x}^{(i)}$
   (b) Update the weights according to:
   $$w_j(k+l) = w_j(k) + \alpha(d^{(i)} - z^{(i)})x_j^{(i)} \tag{16}$$
   where α is a small constant

14.2 Neural networks

The perceptron can only compute linear classification boundaries! How can we compute nonlinear boundaries?

14.2.1 Model for an individual neuron

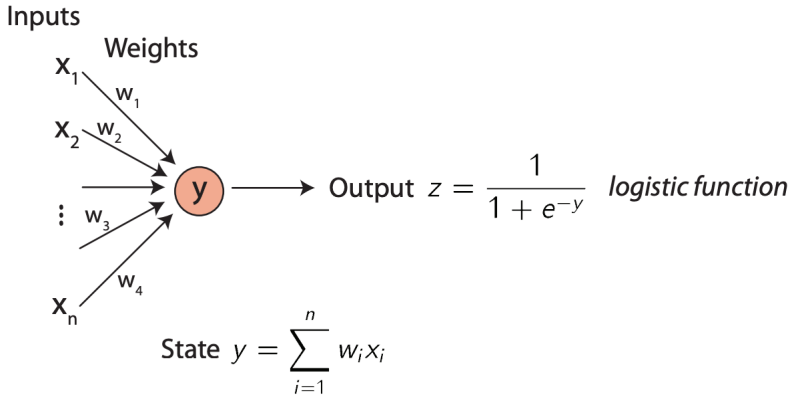

Figure 23: Neural network: model for an individual neuron

14.2.2 Connecting neurons into a network

Multiple layers of coupled logistic neurons form a neural network:

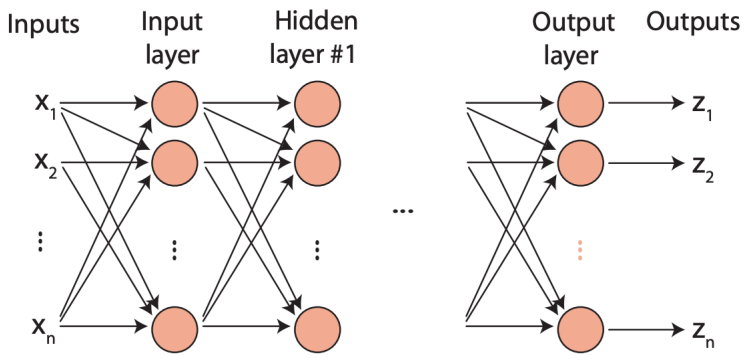

Figure 24: Multiple layers of coupled logistic neurons

14.2.3 Training

As for the perceptron, many successive training steps:

- Forward propagation: feed training data $\mathbf{x}^{(i)}$ to the network to create output $z^{(i)}$
- Backward propagation: propagate target pattern back trough the network to create "desired" output values for each neuron

To train the weights, we update them iteratively using gradient descent. This loss function we want to minimise is:

$$L = \frac{1}{2} \sum_{j \in L_m} (t_j - o_j)^2 \tag{17}$$

where t is the "target" and o is the actual output of that neuron. The weights are updated iteratively as per:

$$w_{kj} = w_{kj} - \alpha \frac{\partial L}{\partial w_{kj}} \tag{18}$$

Here α is called the "learning rate".

However, how do we compute the gradient of the loss function? This can be done using the Backpropogation algorithm, where the gradient with respect to weights in layer L can be computed by using gradients with respect to outputs in layer $L + 1$.

14.3 Problems with Neural Networks

- Convergence to a local minimum (instead of global minimum)
- Slow convergence
- Overfitting (too many neurons/layers) - this leads to poor performance on test data

15 Auditory models and machine hearing

15.1 Gammatone filterbank

The gammatone filterbank decomposes a signal by passing it through a bank of gammatone filters equally spaced on the ERB scale. Gammatone filter banks were designed to model the human auditory system.

The cochlea can be imagined as a series of parallel oscillating strings. The oscillation of each string can be pragmatically modelled with a gammatone function:

$$x(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi f_0 t + \phi) \tag{19}$$

Where a is the amplitude, n is the filter order, b is the filter bandwidth, and f_0 is the center frequency.

15.1.1 Transmission

When we have a good fit using a gammatone, we can look at the transmission. We consider not just a force at time T_0, but a force which oscillates at frequency f (which is principle can be different from f_0). We can then ask what is the steady state amplitude that this steady state force causes?

Given $F = \tilde{F}(f)\cos(2\pi ft)$, an amplitude of $|\tilde{x}(t)|$ will be evoked. This defines the transmission coefficient, which is the ratio of the evoked amplitude and to amplitude of the applied force.

$$\text{Transmission coefficient} = \frac{|\tilde{x}(f)|}{|\tilde{F}(f)|} \tag{20}$$

The Transmission coefficient is a bell shaped curved centred at f_0 (most transmission at the resonance frequency, as expected), with bandwidth b. The faster the exponential decay of the gammatone, the narrower the bandwidth of the transmission coefficient.

15.1.2 Modelling a cochlea

An array of gammatone filters is a "gammatone filterbank" and this can be used to model the cochlea.

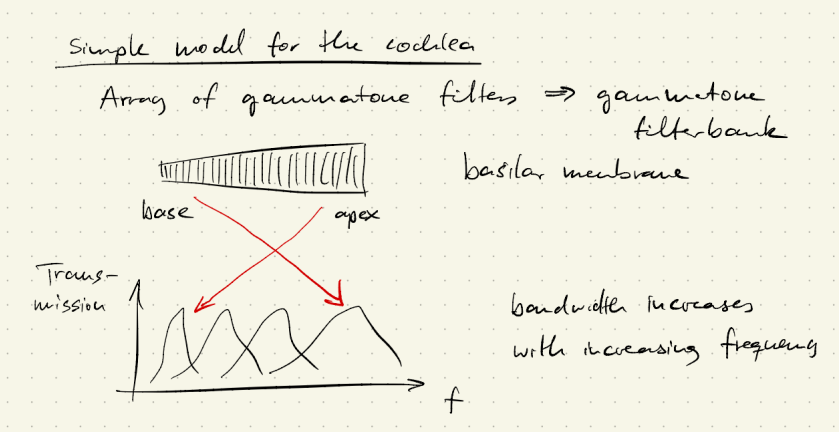

Figure 25: Modelling the cochlea with a gammatone filterbank

15.1.3 Problems with gammatones as a model for the cochlea:

- Gammatone filters are roughly symmetric around the center frequency whereas each cochlear filter is strongly asymmetric
- Gammatone filters are linear in sound intensity, whereas the cochlea has a strong compressive nonlinearity

15.1.4 Cochleagram

Gammatone filters can be used to define a cochleagram. The cochleagram summarises the firing pattern along the basilar membrane in time.

In the gammatone, we have both positive and negative excursions. However, the firing rate can never be below 0. Therefore, we do a half wave rectification, keeping only the positive parts of the signal. This means that the firing rate still reflects the periodicity in the basilar membrane motion and has the same peaks.

15.2 Autocorrelation

Firing in the auditory nerve affects phase locking (which may be useful for machine hearing), so it is useful to extract it from the cochlea channels. The firing rate $f(t)$ can be extracted using autocorrelation.

15.2.1 Short time autocorrelation

$$g(t,\tau) = \int_0^\infty f(t-u)f(t-\tau-u)w(u)du \tag{21}$$

where $w(u)$ is a windowing function.

The windowing function is used because for any speech signal, the periodicity is not constant.

If $f(t)$ is periodic, this will be visible in the short time autocorrelation, with a peak at T, a smaller peak at $2T$ etc.

However, there are problems when the period T changes over time. Therefore, we use triggering.

15.2.2 Triggered autocorrelation

Divide time series of half wave rectified waveform into multiple sections and then layer them on top of each other. When this is done, the peaks do not align. To use trigerring, in each segment you look for the highest peak, and then use these peaks to align the waveforms with each other. This makes the waveforms roughly align.

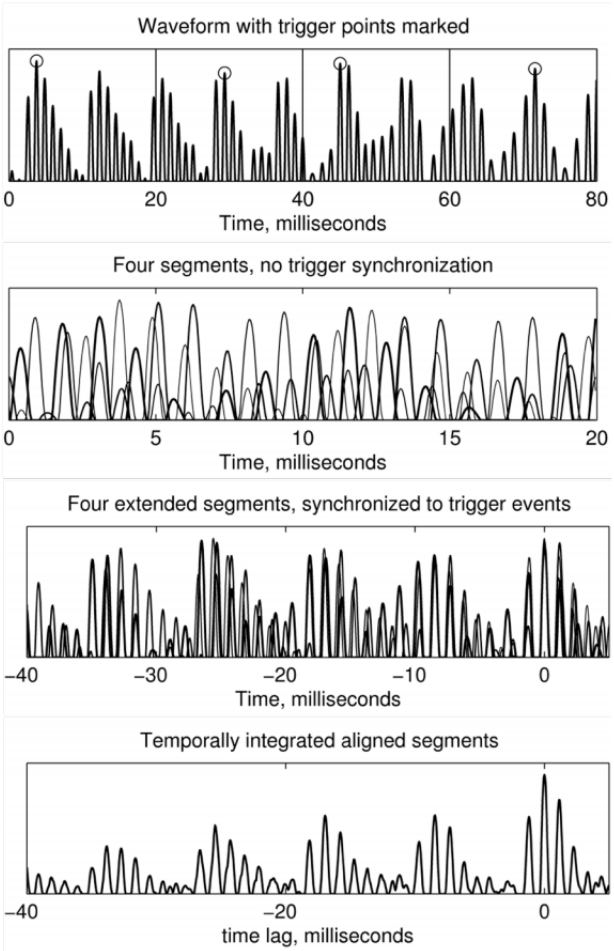

Figure 26: Triggered autocorrelation

This can be done to a cochleagram to obtain a stabilized auditory image.

SAIs can be used to perform "Sound Search", often more successfully that MFCCs.

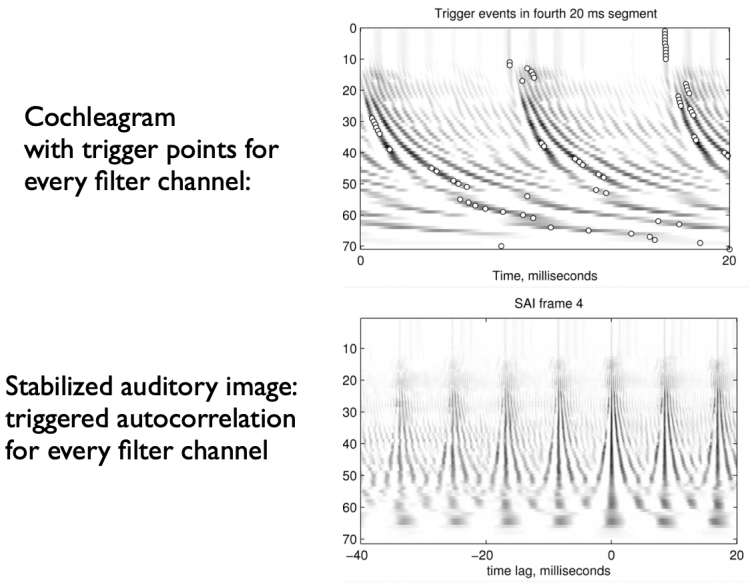

Figure 27: Stabilized auditory image

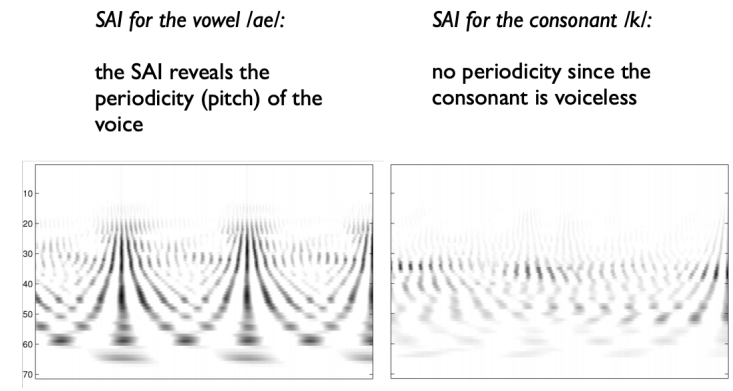

Figure 28: Stabilized auditory image for a vowel and consonant

16 Cochlear implants

A lot of hearing loss is sensorineural, which means it affects the mechanotransduction in the inner ear.

Cochlear implants aim to restore hearing by bypassing the natural mechanotranduction stage by directly producing electrical signals from captured sound in order to stimulate the auditory nerve fibres appropriately.

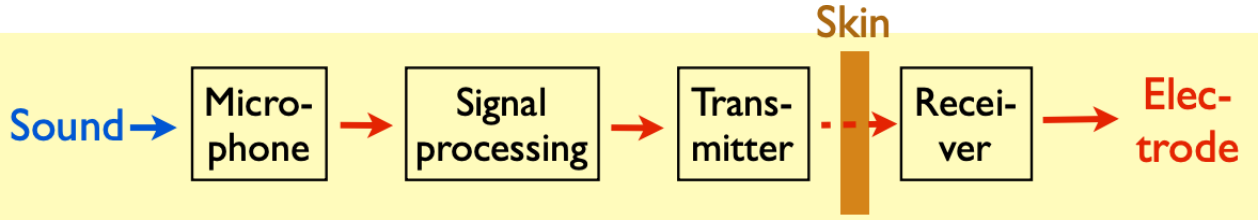

Figure 29: Cochlear implant parts

16.1 Evolution of cochlear implants

16.1.1 Around 1980: F0/F2 processor (Australia)

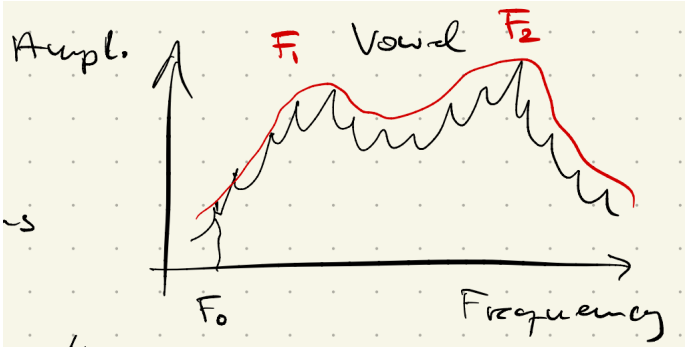

Figure 30: Formants and fundamental frequency of vowel sound

Idea was to find the fundamental frequency F_0 and second formant F_2 and represent them as follows:

- Stimulate electrode at a location corresponding to F_2
- Stimulate at rate F_2 of pulses
- Later: add F_1

16.1.2 Early 1980s: Compressed Analogue Stimulation (USA)

Multiple band pass filters in parallel followed by nonlinear compression. Outputs simulate channels simultaneously. Nonlinear compression means a larger gain for quieter sounds. Also may be called "Automatic Gain Control".

One problem with multi channel simulation is cross talk between the electrodes. Close by electrodes may leak into the domain of another electrodes. This means we cannot have that many electrodes on the cochlear implant.

16.1.3 Early 90s

Continuous Interleaved Sampling: A rapid and non-simultaneous sweep of pulses across the electrode array.

"**n of m approach**" where $n < m$. Stimulate n channels of the available m where we have the largest amplitude.

16.1.4 Current research

Increase spectral resolution, e.g. current steering. Apply a current to each electrode that is different in nature. This allows us to "mix" currents, make it seem we are in between channels.

16.2 Temporal information in neural response: phase locking

Temporal information is important for:

- pitch perception
- frequency discrimination
- music perception
- differentiating between two competing speakers

However, temporal information is not conveyed in standard stimulation strategies in cochlear implants.

16.2.1 Harmonic Single-Band Encoder

1. Extract fundamental frequency (F_0) and harmonics
2. Select largest harmonics and identify frequency-matched electrodes
3. Compress and stimulate electrodes with pulses modulated by a rectified sinusoid at frequency F_0, at a strength of the corresponding harmonics

This can give some encouraging results to improve pitch perception for users of cochlear implants.

17 Interesting research

Shannon and his colleagues found that speech can still be understood even when spectral information is greatly reduced (Shannon et al., Science 1995). In particular, they degraded speech by reducing it to the temporal modulation in only a few frequency bands. According to their results, the smallest number of frequency bands that allowed humans to still identify vowels correctly (with a 90% accuracy rate), was 3.

Tremblay et al. (Nature 2003) investigated the somatosensory basis of speech production. In particular, they altered the jaw movement of human subjects by attaching a complex mechanical load to the jaw. They trained subjects in speaking with the attached load, and then detached the load again. The training was done for three tasks: vocalized speech, silent speech and non-speech movements. The jaw motions differed for vocalized speech and for silent speech, but not for nonspeech movements.