Lecture 6: Frequency Scales: Semitones, Mels, and ERBs

Mark Hasegawa-Johnson

ECE 417: Multimedia Signal Processing, Fall 2020

Outline

Fletcher's Model of Masking, Reviewed

1. The human ear pre-processes the audio using a bank of bandpass filters.
2. The power of the noise signal, in the filter centered at f_c, is

$$N_{f_c} = 2\int_0^{F_S/2} R(f)|H_{f_c}(f)|^2 df$$

3. The power of the tone is $T_{f_c} = A^2/2$, if the tone is at frequency f_c.
4. If there is any band in which

$$10\log_{10}\left(\frac{N_{f_c}+T_{f_c}}{N_{f_c}}\right) > 1\text{dB}$$

then the tone is audible. Otherwise, not.

Equivalent rectangular bandwidth (ERB)

The frequency resolution of your ear is better at low frequencies. In fact, the dependence is roughly linear (Glasberg and Moore, 1990):

$$b \approx 0.108f + 24.7$$

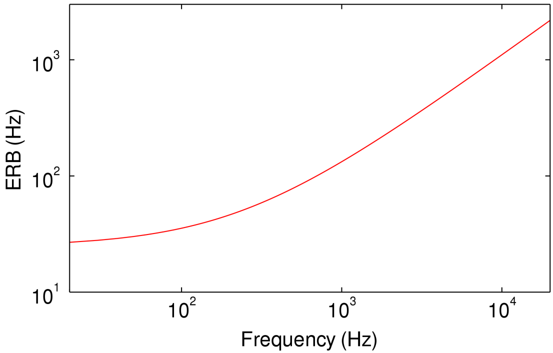
These are often called (approximately) constant-Q filters, because the quality factor is

$$Q = \frac{f}{b} \approx 9.26$$

The dependence of b on f is not quite linear. A more precise formula is given in (Moore and Glasberg, 1983) as:

$$b = 6.23\left(\frac{f}{1000}\right)^2 + 93.39\left(\frac{f}{1000}\right) + 28.52$$

Equivalent rectangular bandwidth (ERB)



By Dick Lyon, public domain image 2009, `https://commons.wikimedia.org/wiki/File:ERB_vs_frequency.svg`

The Gammatone Filter

The shape of the filter is not actually rectangular. Patterson showed that it is

$$|H(f)|^2 = \frac{1}{(b^2 + (f - f_0)^2)^4}$$

He suggested modeling it as

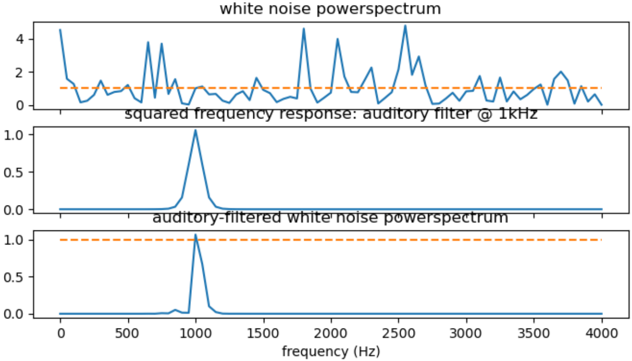
$$H(f) = \left(\frac{1}{b + j(f - f_0)}\right)^n + \left(\frac{1}{b + j(f + f_0)}\right)^n$$

Whose inverse transform is a filter called a **gammatone filter**.

$$h(t) \propto t^{n-1} e^{-2\pi bt} \cos(2\pi f_0 t) u(t)$$

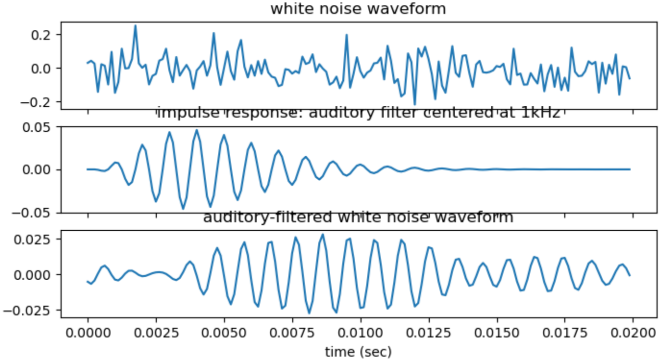
The Gammatone Filter: Spectrum

The top frame is a white noise, $x[n]$. The middle frame is a gammatone filter at $f_c = 1000$Hz, with a bandwidth of $b = 128$Hz. The bottom frame is the filtered noise $y[n]$.



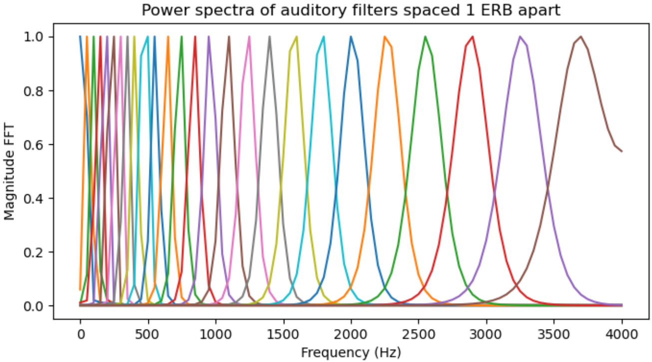
The Gammatone Filter: Impulse Response

The top frame is a white noise, $x[n]$. The middle frame is a gammatone filter at $f_c = 1000$Hz, with a bandwidth of $b = 128$Hz. The bottom frame is the filtered noise $y[n]$.



Frequency responses of the auditory filters

Here are the squared magnitude frequency responses ($|H(\omega)|^2$) of 26 of the 30000 auditory filters. I plotted these using the parametric model published by Patterson in 1974:



Implications for Speech and Audio Recognition

- Different human words must be audibly different.
- If a human being can't hear the difference, then they can't be different words.
- If humans can't hear small differences at high frequencies, then those differences can't possibly change the meaning of a word.
- For speech recognition, we should represent the low-frequency spectrum as precisely as the human ear represents it, and the high-frequency spectrum as imprecisely as the human ear represents it.

Outline

Musical Pitch

Muse playing the lyre, by the Achilles Painter. Staatliche Antikensammlungen, public domain image by Bibi Saint-Pol, `https://commons.wikimedia.org/wiki/File:Mousai_Helikon_Staatliche_Antikensammlungen_Schoen80_n1.jpg`

Pythagorean Tuning

- Humans have always known that $f_2 = 2f_1$ (length of one string is twice the length of the other) means they are an octave apart ("same note").
- A 3:2 ratio ($f_2 = 1.5f_1$) is a musical perfect fifth.
- Pythagoras is attributed with a system of tuning that created an 8-note scale by combining 3:2 and 2:1 ratios ("Pythagorean tuning"), used in some places until 1600.

Equal-Tempered Tuning

Equal-tempered tuning divides the octave into twelve equal ratios.

- **Semitones:** the number of semitones, s, separating two tones f_2 and f_1 is given by

$$s = 12\log_2\left(\frac{f_2}{f_1}\right)$$

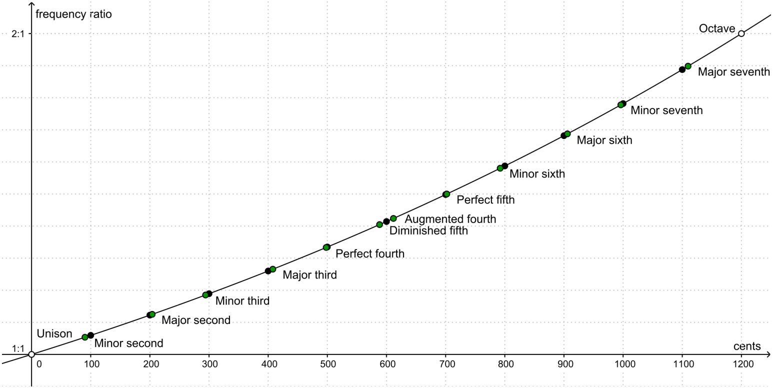
- **Cents:** the number of cents, n, separating two tones f_2 and f_1 is given by

$$n = 1200\log_2\left(\frac{f_2}{f_1}\right)$$

Pythagorean vs. Equal-Tempered Tuning

$\boxed{\text{Pythagorean, Equal-Tempered, and Just Intonation}}$

Pythagorean vs. Equal-Tempered Tuning



By SharkD, public domain image, `https://commons.wikimedia.org/wiki/File:Music_intervals_frequency_ratio_equal_tempered_pythagorean_comparison.svg`

Acoustic Features on a Semitone Scale: the Constant-Q Transform

- Gautham J. Mysore and Paris Smaragdis, Relative pitch estimation of multiple instruments, ICASSP 2009
- Christian Schörkhuber, Anssi Klapuri and Alois Sontacchi, Audio Pitch Shifting Using the Constant-Q Transform, Journal of the Audio Engineering Society 61(7/8):562-572, 2013
- Massimiliano Todisco, Héctor Delgado and Nicholas Evans, A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients, Speaker Odyssey 2016, pp. 283-290

Constant-Q transform

Just like an STFT, suppose that we want our features to be

$$X[k,n] = x[n] * h_k[n]$$

but now suppose we want the filters to be spaced exactly one semitone apart, starting at note A1 on the piano:

$$f_k = 55 \times (2)^{k/12}$$

and suppose that, instead of every filter having the same bandwidth, suppose we want the bandwidth, b_k, to be exactly one tone (one sixth of an octave). That means we want the quality factor to be constant:

$$Q = \frac{f_k}{b_k} = 6$$

Bandwidth of Rectangular & Hamming windows

The rectangular window is

$$w_R[n] = u[n] - u[n-N] \leftrightarrow W_R(\omega) = e^{-j\omega\frac{N-1}{2}} \frac{\sin(\omega N/2)}{\sin(\omega/2)}$$

The Hamming window is

$$w_H[n] = \left(0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right)\right) w_R[n]$$

We can estimate bandwidth by finding the frequency of the first null. That would give

$$\textbf{Rectangular}: \quad b = \frac{F_s}{N}\text{Hz} = \frac{2\pi}{N}\frac{\text{radians}}{\text{sample}}$$

$$\textbf{Hamming}: \quad b = \frac{2F_s}{N}\text{Hz} = \frac{4\pi}{N}\frac{\text{radians}}{\text{sample}}$$

Constant-Q transform

Putting it all together, we get the "Constant Q Transform" as:

$$X_{CQT}[k,m] = x[n] * h_k[-n], \quad h_k[n] = w_k[n]e^{j\omega_k n}$$

where $w_k[n]$ is a window with a length given by

$$Q = \frac{f_k}{b_k}, \quad b_k = \frac{F_s}{N[k]} \quad \Rightarrow N[k] = \frac{F_s}{f_k}Q$$

Outline

Geometric pitch or perceived pitch?

- Hertz: tones are equally spaced on a linear scale
- Semitones: tones are equally spaced on a logarithmic scale
- Mel: equally-spaced tones sound equally far apart, regardless of how high or low the pitch.

A Scale for the Measurement of the Psychological Magnitude Pitch

John E. Volkmann and Stanley S. Stevens

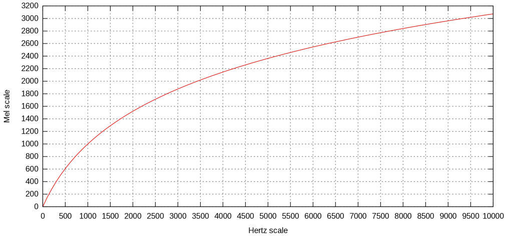
Volkmann and Stevens used the following procedure:

- Play listeners a sequence of three notes, e.g., Play
- Ask the listener to choose a fourth note such that Note4-Note3 = Note2-Note1
- Define the mel-scale (short for "melody") such that mel(Note4)-mel(Note3) = mel(Note2)-mel(Note1)

The Mel Scale

Result: the Mel scale is roughly linear at low frequencies, roughly logarithmic at high frequencies.

$$\text{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$



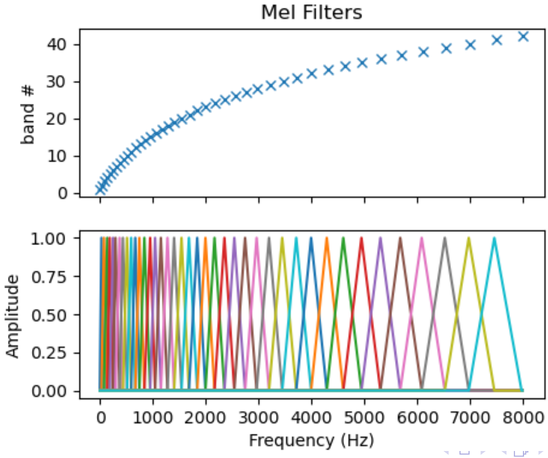
By Krishna Vedala, GFDL, `https://commons.wikimedia.org/wiki/File:Mel-Hz_plot.svg`

Acoustic Features on a Perceptual Scale: Mel-Frequency Cepstrum and Filterbank

- **MFCC**: Steven Davis and Paul Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. ASSP 28(4):357-366, 1980
- **Filterbank Coefficients:** Jinyu Li, Dong Yu, Jui-Ting Huang and Yifan Gong, Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM, 2012 IEEE SLT
- Beth Logan, Mel-Frequency Cepstral Coefficients for Music Modeling, ISMIR 2000

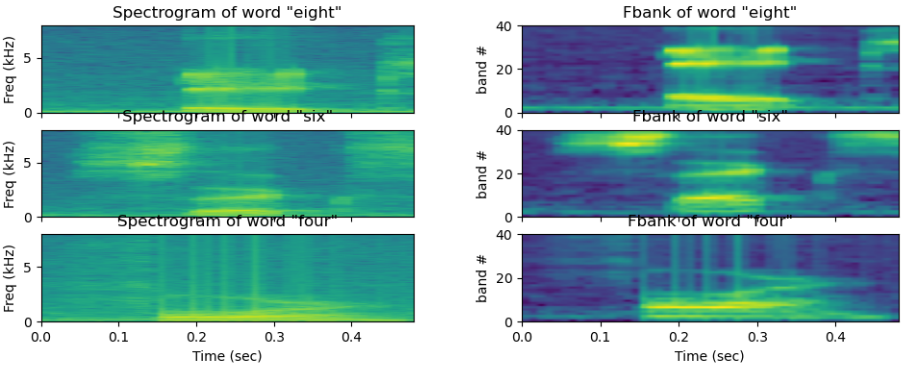
Mel Frequency Filterbank Coefficients

Suppose $X[k,m]$ is the STFT. The mel filterbank coefficients are $C[\ell,m] = \ln\left(\sum_k w_\ell[k]|X[k,m]|\right)$, where the weights, $w_\ell[k]$, are triangles:



Mel vs. Linear Frequency

Mel-frequency (on the right) stretches out the low frequencies more:



Outline

Four frequency scales

- Hertz: tones are equally spaced on a linear scale
- Semitones: tones are equally spaced on a logarithmic scale
- Mel: equally-spaced tones sound equally far apart, regardless of how high or low the pitch.
- ERBs: tones are $<$ 1ERB apart iff their auditory filters overlap, $\geq$ 1ERB apart otherwise.

ERB scale

Let b be the equivalent rectangular bandwidth of a filter centered at f. We want to design a transform $e(f)$ so that

$$e(f+b) \approx e(f) + 1$$

$$e(f+b) - e(f) \approx 1$$

$$\frac{e(f+b) - e(f)}{b} \approx \frac{1}{b}$$

$$\frac{de}{df} = \frac{1}{b}$$

ERB scale

Using the linear approximation,

$$\frac{de}{df} = \frac{1}{0.108f + 24.7}$$

we get

$$e(f) = \frac{1}{0.108} \ln\left(0.108f + 24.7\right)$$

- It looks a lot like the mel scale! Linear at low frequencies, logarithmic at high frequencies.
- ERBs cut over from linear scale to log-scale at $f = \frac{24.7}{0.108} = 228$Hz, vs. Mels, which cut over at $f = 700$Hz.
- The scale is $\frac{1}{0.108} = 9.2$ERBs/octave: smaller than a tone, larger than a semitone.

ERB scale

Using the quadratic approximation,

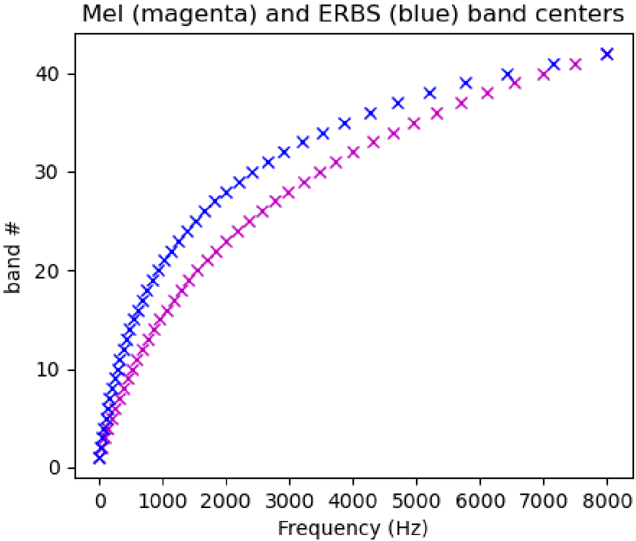
$$\frac{de}{df} = \frac{1}{6.23\left(\frac{f}{1000}\right)^2 + 93.39\left(\frac{f}{1000}\right) + 28.52}$$

we get

$$e(f) = 11.17268 \ln\left(1 + \frac{46.06538 f}{f + 14678.49}\right)$$

- Still linear at low frequencies, logarithmic at high frequencies.
- Linear-to-log cutover is $\frac{14678}{46} = 319$Hz.
- 11.1 ERBs/octave: very close to a semitone

ERBs vs. Mel


Mel (magenta) and ERBS (blue) band centers
band #
0
10
20
30
40
0
1000
2000
3000
4000
5000
6000
7000
8000
Frequency (Hz)

Gammatone filterbank coefficients

Gammatone filterbank coefficients are computed as

$$X[k,m] = x[n] * h_k[n]$$

where we usually use Patterson's gammatone filterbanks:

$$h_k(t) = t^3 e^{-2\pi b_k t} \cos(2\pi f_k t) u[n]$$

Spaced equally on an ERBS scale:

$$f_k = \mathrm{ERB}^{-1}(k)$$

Gammatone filterbank coefficients for speech and audio

- Many papers have had slightly better results using gammatone filterbank coefficients instead of mel filterbank coefficients, at the cost of greater computational cost (because gammatone coefficients are $\mathcal{O}\left\{N^2\right\}$, while MFFB are $\mathcal{O}\left\{N\log_2(N)\right\}$.
- Many recent papers since 2019 use learned filterbank coefficients, instead of gammatone filterbank coefficients.

Outline

Summary

- Semitones and Constant-Q filterbank:

$$f_k = 55(2)^{k/12}$$

$$X[k,n] = x[n] * w_k[n]e^{j\omega_k n}$$

- Mel-Frequency Filterbank coefficients:

$$\text{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

$$C[\ell, m] = \ln\left(\sum_k w_\ell[k]|X[k,m]|\right)$$

- ERBS and Gammatone filterbank:

$$e(f) = 11.17268 \ln\left(\frac{1 + 46.06538f}{f + 14678.49}\right)$$

$$X[k,n] = x[n] * \left(t^3 e^{-2\pi b_k t}\cos(2\pi f_k t)u[n]\right)$$