Lecture Notes in Speech Production, Speech Coding, and Speech Recognition

Mark Hasegawa-Johnson
University of Illinois at Urbana-Champaign

February 17, 2000

Chapter 6

Engineering Models of Audition

6.1 Models of Hearing

An acoustic signal $s(t)$ is transduced from acoustic energy to neural impulses by roughly the following process:

1. The stiffness and mass of the basilar membrane vary continuously from one end of the cochlea to the other. As a result, each place x_i along the basilar membrane is tuned to a specific input frequency ω_i, and vibrates with high amplitude when the input $s(t)$ contains energy at frequencies near ω_i. The mechanical frequency selectivity of the basilar membrane can be modeled using a bank of (relatively broad) bandpass filters, with center frequencies ω_i. The vibration of the basilar membrane at position x_i is then
$$s_i(t) = s(t) * h_i(t) \tag{6.1}$$
where $h_i(t)$ is the impulse response of the ith bandpass filter.

2. Motion of the basilar membrane is transduced by the bending of tiny hair cells implanted in the membrane. These hair cells only detect motion of the membrane in one direction, so in a sense, they half-wave rectify the signal:
$$y_i(t) = \begin{cases} s_i(t) & s_i(t) > 0 \\ 0 & s_i(t) < 0 \end{cases} \tag{6.2}$$

3. The base of each hair cell is innervated with peripheral auditory nerves. These nerves fire all of the time, but they tend to fire more when the hair cell bends. It is useful to talk about the average firing rate $r(t, \omega_i)$ of nerves at a given place x_i along the cochlea. The firing rate $r(t, \omega_i)$ increases monotonically but not linearly with the amplitude of the hair cell bending. The relationship between signal amplitude and firing rate is sometimes approximated by a power law, with $\alpha < 1$:
$$r(t, \omega_i) \approx y_i(t)^\alpha \tag{6.3}$$

So far, we have sketched the following model of auditory processing. This model is not complete, but most complete models include at least the blocks shown here:

$$s(t) \rightarrow \boxed{\text{BPF at } \omega_i} \rightarrow \boxed{\text{HW Rectify}} \rightarrow \boxed{\text{Power-Law Compress}} \rightarrow r(t, \omega_i) \tag{6.4}$$

6.2 Perceptual Frequency Scales

One of the first requirements for a perceptually based ASR front end is a scale for the BPF center frequencies. Ideally, we would like some transformation from perceptual frequencies to physical frequencies, so that the ω_i can be uniformly distributed on a perceptual scale:

$$\omega_i = \theta(b_i), \quad b_i = 1, 2, \ldots \tag{6.5}$$

Fortunately, many different types of experiments point to the same type of function $\theta(b_i)$.

6.2.1 Critical-Band Scale

Remember that in the auditory model, amplitude compression takes place after the bandpass filter. Suppose a sound is made up of the sum of two sinusoids,

$$x(t) = A_1 \cos(j2\pi f_1 t) + A_2 \cos(j2\pi f_2 t) \tag{6.6}$$

If f_1 and f_2 are so close that they fit within the same BPF, the output nerve firing rate is:

$$r(t, x_i) \sim (A_1 + A_2)^\alpha \tag{6.7}$$

If f_1 and f_2 go through different BPFs, the sum of the nerve firing rates is:

$$\sum_i r(t, x_i) \sim A_1^\alpha + A_2^\alpha > (A_1 + A_2)^\alpha \tag{6.8}$$

By this logic, $x(t)$ will sound louder if f_1 and f_2 are far apart than it will if they are close together. This is, in fact, what happens. The distance $|f_2 - f_1|$ at which $x(t)$ begins to sound louder is called a "critical bandwidth." At $f = 1000$Hz, a critical band is 160Hz wide; at higher and lower bandwidths, the bandwidth is a function of frequency. The frequency scale given by critical bandwidths is called the Bark scale, $b(2\pi f)$:

$$b(2\pi f) \approx \begin{cases} \frac{f}{100} & f < 600\text{Hz} \\ 6 \log f + \text{constant} & f > 1500\text{Hz} \end{cases} \tag{6.9}$$

One bark equals the width of one critical band centered at $\omega_i = \theta(b)$.

6.2.2 Mel-Frequency Scale

The mel-frequency scale is characterized as follows:

- 1000 mels = 1000 Hz.
- Experimental subjects will tell you that a tone at $x/2$ mels is "half as high" as a tone at x mels.

It turns out that the Bark scale and mel scales are proportional to each other; 1 bark is approximately equal to 100 mels.

6.2.3 Place on the Basilar Membrane

By studying the response of the basilar membrane to sinusoids of different frequencies, it is possible to map out a relationship between place x_i and filter center frequency ω_i. Based on these experiments, it can be shown that a frequency spacing of one Bark corresponds to a separation along the basilar membrane of about 1.2mm.

6.2.4 Frequency Just-Noticeable-Difference

Unlike the three experiments just cited, the ability of listeners to differentiate two similar frequencies does *not* follow a Bark scale.

Pitch Discrimination

The ability of listeners to discriminate between two different pitches is dependent on

- The frequency at which the tones are played.
- The type of tone (pure tone, or complex tone e.g. speech).
- The loudness of the tone (pitch discrimination is harder with quiet signals).

With pure tones at comfortable listening levels between 500Hz and 2000Hz, listeners can distinguish tones separated by about 3Hz. At frequencies higher than 2000Hz, the just-noticeable-difference (JND) Δf rises faster than f.

Formant Frequency Discrimination

Formant discrimination depends on

- The formant frequency.
- The loudness of the signal.
- The distance between the formant frequency and the nearest pitch harmonic.

Listeners notice formant shifts most easily when the formant switches affiliation from one pitch harmonic to a different pitch harmonic. Formant shifts in the vicinity of a pitch harmonic are harder to hear. However, as a general rule of thumb, formant discrimination in the F1-F3 range can be approximately modeled using Weber's law:

$$\frac{\Delta F_n}{F_n} \approx \text{constant} \tag{6.10}$$

where the constant is about 3-5%.

6.3 Loudness

The perceived loudness of a tone depends on

- Intensity (sound pressure level).
- Frequency.
- Manner of presentation: alone, or with other signals that might mask the tone.

6.3.1 Loudness vs. Frequency

Figure 6.3.1 shows an example of an equal-loudness contour, which is a function of frequency. Thus, for example, a 100Hz tone with an sound pressure level of about 1 dyne/cm^2 sounds about as loud to most listeners as a 1000 Hz tone at about 0.1 dyne/cm^2. Both of these tones have a "loudness" of 40 phons, which means that the 1000Hz tone is about 40dB more intense than the minimum audible tone at 1000Hz.

- The absolute threshold of hearing is lowest between 500Hz and 4000Hz, that is, in the approximate range of the lower formant frequencies.
- Above the absolute threshold level, loudness increases most rapidly as a function of sound pressure level at very low and very high frequencies.
- The equal-loudness curves tend to flatten out at higher sensation levels, for frequencies below about 10kHz.

6.3.2 Intensity JND

Just-noticeable differences for a pure tone depend on listening level, and at low intensities the JND is also a function of frequency. At comfortable listening levels (about 60phon), the intensity JND, ΔI, varies approximately according to Weber's law:

$$\frac{\Delta I}{I} \approx c_1 \tag{6.11}$$

$$10 \log_{10} \frac{I + \Delta I}{I} \approx c_2 \tag{6.12}$$

…for some constants c_1 and c_2. For pure tones, JND thresholds as low as $c_2 = 0.6$dB have been reported. For more complex sounds, like speech, the overall intensity JND is higher (about 1.5dB). A good rule of thumb to remember is that at comfortable listening levels, for average sounds,

$$10 \log_{10} \frac{I + \Delta I}{I} \approx 1\text{dB} \tag{6.13}$$

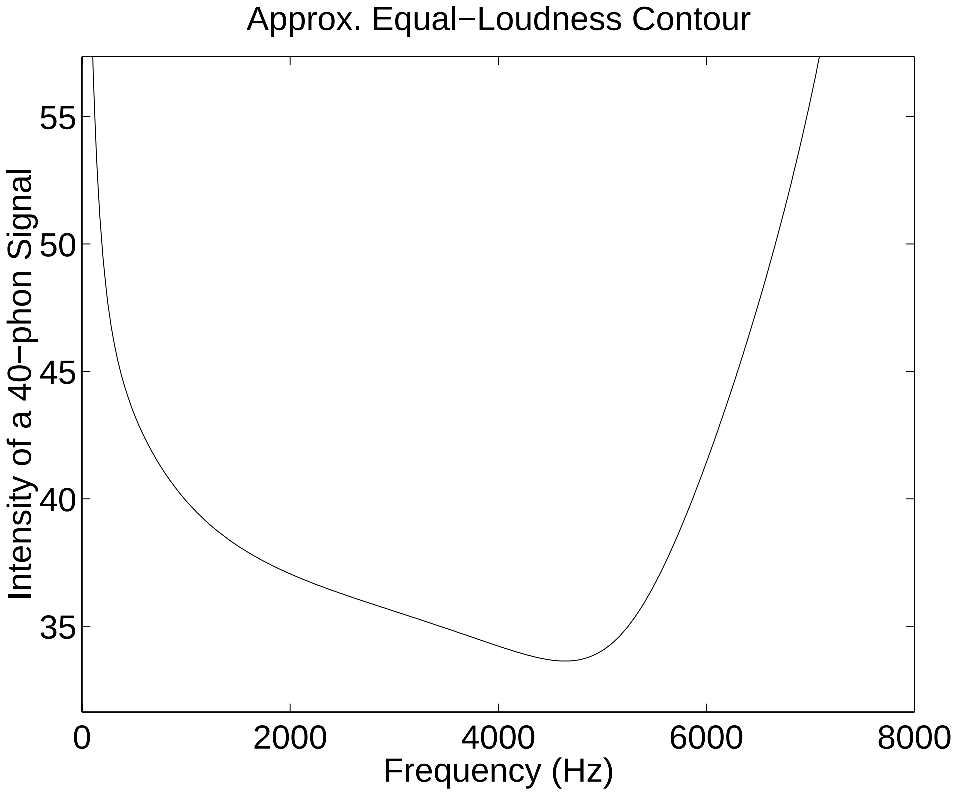

Figure 6.1: An example of an equal-loudness curve as a function of frequency.

6.4 Masking

Remember that loudness is integrated by the bandpass filters in the cochlea. It is possible to measure the shape of the bandpass filters through masking experiments, which proceed roughly as follows:

1. Play the subject a "masking signal" $X(\omega)$.
2. Play the same signal, plus a probe tone at relatively low level.
3. Ask the subject which of the two signals contained a probe tone (forced choice).

If subjects can not correctly guess which signal contained the probe tone, then we can conclude that the probe $P(\omega)$ was "masked" by the louder signal $X(\omega)$. By repeating this experiment with several different probe and masker signals, it is possible to map out the shape of the critical band spectrum $H_i(\omega)$. Roughly, $H_i(\omega)$ must be shaped so that:

$$\frac{\Delta I}{I} = \frac{\int |H_i(\omega)P(\omega)|^2 d\omega}{\int |H_i(\omega)X(\omega)|^2 d\omega} \begin{cases} < \text{threshold} & \text{Probe was masked} \\ > \text{threshold} & \text{Probe was not masked} \end{cases} \tag{6.14}$$

It turns out that the filters are nonlinear: as the level of the masker increases, the filter bandwidth broadens. At moderate listening levels, the filter shapes are as shown in R&J Fig. 3.50. Note that:

- Filters overlap.
- Filter tails drop off more slowly toward negative than positive frequency, meaning that LF sounds mask HF sounds more than HF sounds mask LF.
- The filters above about 400Hz have filter characteristics which are similar (though not quite the same) on a Bark-frequency scale:

$$H_i(b) \approx H_0(b - b_i) \tag{6.15}$$

where $H_0(b)$ is a prototype critical band filter, and b_i is the center frequency of the ith filter (in Bark).

6.4.1 Noise-Masker Ratio

By now, it should be clear that much of the information in an acoustic signal is not even audible. In audio coding, it is possible to add noise which is totally inaudible if we make sure that the noise being added is totally masked by the recorded audio signal.

Define the quantization noise spectrum to be $Q(b)$, and the noise within a single critical band to be

$$N(b_i) = \frac{1}{2B}\int_{-B}^{B} |H_i(\beta)|^2 |Q(\beta)|^2 d\beta \tag{6.16}$$

If we assume that

$$H_i(b) \approx H_0(b - b_i) \tag{6.17}$$

then $N(b)$ is

$$N(b) = \frac{1}{2B}\int_{B}^{B} |H_0(\beta - b)|^2 |Q(\beta)|^2 d\beta = |H_0(-b)|^2 * |Q(b)|^2 \tag{6.18}$$

The noise component $N(b)$ is inaudible if:

1. The noise is below the absolute threshold of hearing, $A(b)$:

$$N(b) < A(b) \tag{6.19}$$

or

2. The noise is masked by a nearby frequency component of the signal $X(b)$:

$$\frac{\Delta I}{I} = \frac{N(b)}{|H_0(-b)|^2 * |X(b)|^2} < \text{threshold} \tag{6.20}$$

Thus, it is possible to define a "masking spectrum" $M(b)$:

$$M(b) = \max\left(A(b), \text{threshold} \times |H_0(-b)|^2 * |X(b)|^2\right) \tag{6.21}$$

such that the quantization noise is completely masked if, for all b,

$$\frac{N(b)}{M(b)} < 1 \tag{6.22}$$

Practical sub-band audio coding systems will often calculate such a masking spectrum, and then attempt to shape the noise spectrum $N(b)$ in order to minimize the noise-masker ratio $N(b)/M(b)$ for all b.

6.5 Mel-Frequency Cepstrum

In many speech recognition front ends, a linear-frequency FFT $X(k)$ is converted into a mel-frequency power spectrum by multiplication with models of the critical band filter spectra.

$$\tilde{S}(b) = \sum_{k=0}^{N-1} |H_b(k)|^2 |X(k)|^2, \qquad b = 1, \ldots, K \tag{6.23}$$

The filter spacing is typically chosen to equal between 1.0 and 1.5 Bark, with overlap of about 50% between filters. The resulting mel-scale spectrum, $\tilde{S}(b)$, can be interpreted as a mel-scale power spectral density, or as the summed energy of a series of bandpass filtered signals $h_b(n)$:

$$\tilde{S}(b) = \sum_{n=0}^{N-1} x_b^2(n), \qquad x_b(n) = x(n) \circledast h_b(n) \tag{6.24}$$

If $x(n)$ is windowed such that $N \geq \text{length}(h_b(n)) + \text{length}(x(n)) - 1$, then $x_b(n) = x(n) * h_b(n)$.

The mel-frequency spectrum has the odd property of being undefined for $b = 0$. For this reason, it is sometimes interpreted as a spectrum sampled at frequencies of

$$\theta_b = \frac{\pi(2b-1)}{2K} \tag{6.25}$$

The mel-frequency cepstrum (MFCC) can be defined as the inverse DFT of the symmetric log-mel-frequency spectrum, with frequencies θ_b shifted as shown above:

$$\begin{aligned}
\tilde{c}(m) &= \frac{1}{2K}\left(\sum_{b=1}^{K} \log \tilde{S}(b) e^{\frac{j\pi m(2b-1)}{2K}} + \sum_{b=-K}^{-1} \log \tilde{S}(-b) e^{\frac{j\pi m(2b+1)}{2K}}\right) \tag{6.26}\\
&= \frac{1}{K}\sum_{b=1}^{K} \log \tilde{S}(b) \cos\left[\frac{\pi m}{K}\left(b - \frac{1}{2}\right)\right] \tag{6.27}\\
& \tag{6.28}
\end{aligned}$$

By Parseval's theorem, the difference between two mel-frequency spectra $\tilde{S}_1(b)$ and $\tilde{S}_2(b)$ can also be calculated in the cepstral domain.

$$\frac{1}{K}\sum_{b=0}^{K} \left|\log \tilde{S}_1(b) - \log \tilde{S}_2(b)\right|^2 = \sum_{m=0}^{K} (\tilde{c}_1(m) - \tilde{c}_2(m))^2 \tag{6.29}$$

The cepstral distortion measure can also be smoothed using a cepstral window, as with a regular cepstral distortion measure.

6.6 Perceptual LPC

In perceptual LPC, the mel-frequency or Bark-frequency spectrum is computed, just as it is in computation of the MFCC:

$$\tilde{S}(b) = \sum_{k=0}^{N-1} |H_b(k)|^2 |X(k)|^2, \qquad b = 1, \ldots, K \tag{6.30}$$

The magnitude-squared bark-frequency spectrum is then divided by a model of the minimum audibility threshold $A(b)$, and compressed using a power-law model of loudness:

$$Y(b) = \left(\frac{\tilde{S}(b)}{A(b)}\right)^{0.33} \tag{6.31}$$

Since the inverse transform of a power spectrum is the autocorrelation $R(m)$, the inverse transform of the loudness spectrum can be viewed as a sort of perceptual autocorrelation, $\tilde{R}(m)$:

$$\tilde{R}(m) = \text{DFT}^{-1}\left\{Y(b)\right\} \tag{6.32}$$

Unlike the MFCC, the PLP calculation requires that the spectral samples $Y(b)$ be located at integer harmonics of the first sample frequency:

$$\theta_b = \frac{2\pi b}{2K} \tag{6.33}$$

If we assume that $Y(0) = 0$ and $Y(2K - b) = Y(b)$, the inverse DFT simplifies as follows:

$$\begin{aligned}
\tilde{R}(m) &= \frac{1}{2K}\sum_{b=0}^{2K} Y(b) e^{\frac{j2\pi bm}{2K}} \tag{6.34}\\
&= \frac{1}{K}\sum_{b=1}^{K} Y(b) \cos\left(\frac{\pi bm}{K}\right) + \frac{(-1)^m}{2K} Y(K) \tag{6.35}\\
& \tag{6.36}
\end{aligned}$$

The perceptual autocorrelation is modeled using an LPC polynomial by solving the normal equations

$$\tilde{R}(m) = \sum_{k=1}^{p} \tilde{a}_k \tilde{R}(|m-k|) \tag{6.37}$$

Finally, the LPC coefficients $\tilde{a}_k$ are used to calculate p LPC-cepstral coefficients,

$$\tilde{c}(m) = \tilde{a}_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) \tilde{c}(k)\tilde{a}_{m-k}, \quad 1 \le m \le p \tag{6.38}$$

The PLP representation has the following characteristics which are useful in an ASR front end:

- Very low order: maximum recognition accuracy is obtained with an LPC order of approximately $p = 5$, which is much less than the order of most other spectral representations. Since PLP requires fewer coefficients, PLP coefficients are more efficient in terms of both storage and recognition complexity than most other spectral representations.
- PLP is relatively insensitive to changes in the spectral tilt caused by, for example, changing the microphone.
- PLP may be less sensitive to speaker variations than most spectral representations.

6.7 Exercises

1. **Perceptual LPC** The article handed out in class ("Perceptual linear predictive (PLP) analysis of speech," by Hynek Hermansky) makes use of the folowing transform from linear frequency f to Bark frequency b:

$$b(2\pi f) = 6\log\left\{f/600 + [(f/600)^2 + 1]^{0.5}\right\} \tag{6.39}$$

   (a) Show that, for $f \gg 600$Hz,

$$b(2\pi f) \approx c_1 \log f + c_2 \tag{6.40}$$

   What are the values of the constants c_1 and c_2?

   (b) Show that, for $f \ll 600$Hz,

$$b(2\pi f) \approx c_3 f \tag{6.41}$$

   What is the value of the constant c_3? Hint: make use of the following Laurent expansion

$$\log(1-x) = -\sum_{r=1}^{\infty} \frac{x^r}{r}, \quad |r| < 1 \tag{6.42}$$

   Assume that, if x is very small, $x^2 \approx 0$.

   (c) In Perceptual LPC, the loudness spectrum $Y(i)$ (called $\Phi(\Omega)$ in the article) is calculated for integer values of i using the following equations, where $X(k)$ is the speech DFT, $H_i(k)$ is the bandpass filter for band number i, and $A(i)$ is a rational model of one of the equal-loudness curves shown in Fig. 7.5 of the lecture notes:

$$\tilde{S}(i) = \sum_{k=0}^{N-1} |H_i(k)|^2 |X(k)|^2, \quad i = 0, \ldots, K \tag{6.43}$$

$$Y(i) = \left(\frac{\tilde{S}(i)}{A(i)}\right)^{0.33} \tag{6.44}$$

   $Y(i)$ is calculated for $i = 1, \ldots, K$, but the value of $Y(0)$ is not calculated, because it is known *a priori*. Based on the above formulas, or based on formulas in the Hermansky article, demonstrate that for any $X(k)$ and any set of filters $H_i(k)$,

$$Y(0) = 0 \tag{6.45}$$

   (d) Suppose that, instead of using the critical-band filters given in the Hermansky article, you have decided to implement Bark-frequency analysis using the following set of Bark-frequency filters $\tilde{H}_i(b)$:

$$\tilde{H}_i(b) = \tilde{H}_0(b - i), \quad i = 1, 2, \ldots, K \tag{6.46}$$

$$\tilde{H}_0(b) = \begin{cases} 1 + b & -1 \le b \le 0 \\ 1 - b & 0 \le b \le 1 \\ 0 & \text{else} \end{cases} \tag{6.47}$$

   Suppose, further, that you have decided to implement critical-band analysis using an N-point DFT, as shown in equation 6.43 above. Find a formula for the linear-frequency filter $H_i(k)$ as a function of i, k, N, the sampling frequency F_s, and the function $b(2\pi f)$. Sketch $H_i(k)$.

2. **Perceptual Frequency Scaling**

   You have been hired to design a bank of filters for the front end of a speech recognizer. The filters $H_i(2\pi f)$ will be ideal bandpass filters with the following form:

$$H_i(2\pi f) = \begin{cases} 1 & f_i < f < f_{i+1} \\ 0 & \text{otherwise} \end{cases} \tag{6.48}$$

Your employers want the band edges f_i to be spaced uniformly along some sort of perceptual frequency scale, but they do not want to use the Bark frequency scale, since it gives too much weight to low frequencies. In fact, they have already decided that the first seven band edges should be at $f_1 = 83.3$, $f_2 = 250$, $f_3 = 416.7$, $f_4 = 583.3$, $f_5 = 750$, $f_6 = 916.7$, and $f_7 = 1083.3$ Hertz.

You decide that the best way to proceed is by creating a new frequency scale called the Meow scale, $m(2\pi f)$, which is proportional to the Bark scale $b(2\pi f)$ at very low and very high frequencies:

$$m(2\pi f) = \begin{cases} \alpha_1 b(2\pi f) & \text{if } f \text{ very small} \\ \alpha_2 b(2\pi f) & \text{if } f \text{ very large} \end{cases} \tag{6.49}$$

After some thought, you realize that a two-part Meow scale of the following form will satisfy equation 6.49 without changing the values of f_1 through f_7 specified by your employers:

$$m(2\pi f_{i+1}) = m(2\pi f_i) + 1, \qquad m(2\pi f) = \begin{cases} m_1(2\pi f) & f \le 916.7 \text{ Hz} \\ m_2(2\pi f) & f \ge 916.7 \text{ Hz} \end{cases} \tag{6.50}$$

What are the functions $m_1(2\pi f)$ and $m_2(2\pi f)$?
Hint: you may find it useful to know that $\log(1083.3/916.7) \approx 1/6$.