Power spectrum model of masking

- **Assumptions:**
  - Only frequencies within the passband of the auditory filter contribute to masking.
  - Detection is based on a single auditory filter, centered on the frequency of the tone.
  - Listeners ignore short-term fluctuations in the noise, and do not rely on phase differences between signal and noise.

**Notched noise method**

- Patterson (1976) estimated auditory filter shapes from the function relating tone threshold to notch width.
- The derived filters have a rounded top and steep skirts, with bandwidths 10-15% of filter center frequency.

**Off-frequency listening**

- Tone detection can be improved by shifting filter center frequency to maximize SNR.

**Simulation of reduced frequency selectivity**

- Derived auditory filter shapes
**Auditory filter shapes as a function of frequency**

- The figure shows a frequency response of a gammatone filter bank.
- The x-axis represents frequency (kHz) and the y-axis represents filter gain (dB).

**Auditory filter shapes as a function of level**

- The figure illustrates the output level (dB) as a function of frequency (Hz).
- The x-axis represents frequency (Hz) and the y-axis represents output level (dB).

**Equivalent Rectangular Bandwidth**

- The equivalent rectangular bandwidth (or ERB) of a filter is the bandwidth of a rectangular filter which has the same power output as that filter, when the input is white noise.

**Cochlear frequency-place map**

- Greenwood (1961) developed a function to relate the characteristic frequency (CF) at each place on the cochlea to the distance (x) of that place from the apex.

**Equivalent Rectangular Bandwidth**

- ERB – equivalent rectangular bandwidth of the estimated auditory filter – about 10-15% of the filter center frequency.

**ERB Scale**

- One ERB unit corresponds to a distance of about 0.89 mm along the basilar membrane.
ERB-rate scale
• The ERB-rate scale is a warped frequency scale modeling changes in the ERB of the auditory filter as a function of frequency.

Excitation patterns
• Auditory excitation patterns show the composite output of a bank of simulated auditory filters as a function of filter center frequency.

Excitation patterns
• Excitation patterns provide a good model of auditory frequency selectivity and masking: frequency components that are resolved by the auditory system produce distinct peaks in the excitation pattern.
Simulation studies

- **Simulation** of reduced frequency selectivity (spectral smearing of the short-term speech spectrum) results in lowered intelligibility for listeners with normal hearing, particularly in noise (ter Keurs et al., 1993; Baer & Moore, 1994)

Distortion of spectral shape

- Broader auditory filters produce a “smeared” excitation pattern: reduced prominence of peaks, smaller peak-to-valley ratios.
- Introduction of noise fills up the valleys between the spectral peaks and reduces the distinctiveness of the spectral profile.
Distortion of temporal structure

- Broader auditory filters alter the temporal fine structure of the output.
- Increased contribution of adjacent components
- Increase in within-channel modulation
- Diminished differences between adjacent channels

Effects of reduced frequency selectivity on temporal structure

Loudness Recruitment

- When a sound is increased in level above absolute threshold, the rate of growth of loudness is greater than normal.
- At levels >90-100 dB SPL, loudness returns to normal (sound appears equally loud to hearing-impaired and normal listeners).

Loudness Recruitment

- **Loudness recruitment** is associated with reduced dynamic range (range between absolute threshold and highest comfortable level).
- Recruitment may reduce the ability to “listen in the dips” in a fluctuating masker, such as a competing voice.
- Recruitment distorts loudness relationships among components of speech sounds.

Temporal structure of speech

- Rosen (1992) proposed that the temporal structure of speech can be partitioned into three levels based on their rate of modulation:
  - **Envelope cues** - slow modulations (<50 Hz) associated with syllable structure
  - **Periodicity cues** (70-500 Hz) correspond to the rate of vocal fold vibration (voice pitch)
  - **Fine-structure cues** (> 250 Hz) correspond to rapid modulations associated with formant changes
Temporal modulation structure of speech

- Houtgast and Steeneken (1985) showed that the intelligibility reduction caused by noise and reverberation can be modeled in terms of the corresponding reduction in **temporal envelope modulations**.

Modulation spectrum of speech

- Houtgast and Steeneken proposed a measure called the **Speech Transmission Index** (STI), based on the estimates of the amount of modulation preserved in different frequency bands. The STI is designed to predict the overall intelligibility of distorted speech.

Noise and reverberation tend to fill the dips in the temporal envelope of speech

- Noise and reverberation tend to flatten formant transitions and fill gaps between them.

Modulation spectrum of speech

- The capacity of a communication channel to transmit modulations in the energy envelope is referred to as the temporal **modulation transfer function** (MTF).
- The MTF for speech has a low-pass shape with a peak around 4-6 Hz, reflecting the syllable alternation rate in connected speech.
Signal processing to obtain the MTF

1. Filter the speech signal in octave bands between 0.25 and 8 kHz.
2. Square and low-pass filter the output (30 Hz).
3. Analyze the resulting intensity envelope using 1/3 octave bandpass filters with center frequencies between 0.63 and 12.5 Hz.

Temporal envelope modulations

Channel Vocoder

- Dudley (1939) developed the channel vocoder, a speech analysis-synthesis system that exploits the modulation structure of speech.
- Vocoder belong to a class of speech analysis/synthesis systems that perform a source-filter decomposition of the signal.

Channel Vocoder

- In each channel, the amplitude envelope is extracted from the filtered waveforms.
- A sequence of pulses is generated at the frequency of the fundamental for voiced sounds; or white noise if the signal is unvoiced.
- The envelope is modulated by the pulsed and/or noise source, and summed across channels.

Channel Vocoder

- The channel vocoder filters the speech signal through a bank of bandpass filters with center frequencies distributed across the speech range.

Channel Vocoder

Dudley, JASA 1939

1. Filter speech with a set of bandpass filters.
2. Extract the waveform envelope in each channel.
3. Obtain the excitation signal (pulsed or noise) from the broadband signal.
4. Modulate the excitation signal with the filtered waveform envelope in each channel, then re-filter the result through the same bandpass filter.
5. Sum up the bands and scale to the appropriate amplitude.
Shannon et al. (1995)

- Shannon et al. used a version of the channel vocoder to replace the rich spectrotemporal structure of speech with just four noise bands.
- Their processor eliminates the fine structure of speech (including evidence of voicing and details of spectral shape), but preserves the temporal modulations in four broad frequency channels (Friesen et al., 2001).

Shannon et al. (1995)

- Speech processed through the four-band vocoder remained highly intelligible in quiet (90% or better for vowels in hVd words, consonants in VCV syllables, and words in sentences).
- The findings illustrate the high degree of redundancy in speech and the importance of its temporal modulation structure.
Shannon et al. (1995)

• One interpretation is that the temporal modulations in a small number of frequency bands provide necessary and sufficient information for accurate speech recognition (e.g., Greenberg 1996).
• Are the fine structure of speech and details of spectral shape therefore unnecessary for speech recognition?

Simulation of cochlear implants

• When speech is processed using the algorithm described by Shannon et al. and presented to normal-hearing listeners, the resulting performance levels are similar to cochlear implant users with the same number of electrode channels as filter channels in the noise-excited vocoder (Friesen et al., 2001).

Noise-excited channel vocoder

• Shannon et al. (1995) used a noise-excited channel vocoder to simulate the effects of cochlear implant processing.
  Demo: 16-channel vocoders
• Mixed-excitation channel vocoder
• Noise-excitation channel vocoder

Noise-excited channel vocoder

• Normal-hearing listeners presented with speech processed through a CI simulation experienced about the same degree of difficulty as actual CI users with the same number of electrode channels.

Effects of frequency shifts

• Fu, Nogaki & Galvin (2005) Auditory Training with Spectrally Shifted Speech. JARO 6: 180-189

Eight-channel CI simulation
Stages of processing: Noise carrier

1. Apply a high-pass pre-emphasis filter
2. Filter the signal into 8 spectral bands equally spaced on the Greenwood scale
3. Extract envelope from filtered waveform by half wave rectification and LP filtering at 160 Hz
4. Modulate a sample of broadband noise by the envelope and pass through the same filter
5. Sum the 8 bands and scale to the level of the original signal

Stages of processing: Sinusoidal carrier

- Same, but uses a sinusoid centered at the band's center frequency to generate the carrier, rather than broadband noise.

Stages of processing: Sinusoidal carrier

1. Apply high-pass pre-emphasis filter
2. Filter the signal into 8 spectral bands equally spaced on the Greenwood scale
3. Extract envelope from filtered waveform by half wave rectification and LP filtering at 160 Hz
4. Modulate sinusoidal carrier by the envelope and pass through the same filter
5. Sum the 8 bands and scale to the level of the original signal.

Greenwood scale

Donald D. Greenwood
*A cochlear frequency-position function for several species – 29 years later.*
J. Acoust. Soc. Am. 87 (6), June 1990

- Map of frequency to distance (in mm) along the basilar membrane

Examples

- Original
- 8-channel CI simulation, noise carrier
- 8-channel CI simulation, sinusoidal carrier
Filter bands and carriers

### Table 1

<table>
<thead>
<tr>
<th>Filter Center Frequency (kHz)</th>
<th>Filter Gain (dB)</th>
<th>Carrier bands</th>
<th>Unshifted carriers</th>
<th>Shifted carriers</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>-20</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>0.2</td>
<td>-10</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>60</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>80</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

Sources: Fu et al. (2005) JARO 6: 180–189.

### Examples

- Original
- 8-channel CI simulation, unshifted sinusoidal carrier
- 8-channel CI simulation, shifted sinusoidal carrier

**Conditions**
- 8-channel simulations presented to listeners with normal hearing
- Trained and tested with spectrally shifted speech

<table>
<thead>
<tr>
<th>Test-only protocol</th>
<th>Preview protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targeted vowel contrast training protocol</td>
<td>Sentence training protocol</td>
</tr>
</tbody>
</table>
Results

CI simulations of music
- Instrumental music processed through an 8-channel simulation with noise carriers
- Original

CI simulations of bird song
- Bird song processed through an 8-channel simulation with noise carriers
- Original

Cochlear Implants
- Cochlear implants provide reduced spectral activation – up to 24-channel electrode array replaces large array of mechanical-to-neural transducers (5k inner hair cells).

Cochlear Implants
- Imperfect electrode penetrations cause inappropriate activation of tonotopic map.
Cochlear Implants

- Current spread to adjacent regions leads to **spectral smearing (masking)**.
- Impedance mismatch can lead to imperfect **amplitude mapping** and inappropriate **loudness growth**.

Cochlear Implants and speech perception

- **In quiet**, speech recognition by cochlear implant users is fairly successful (70-80% words correct).
- **Implication**: Loss of spectral resolution (frequency selectivity) has less impact and fewer consequences for speech perception than predicted, suggesting that temporal processing and gross spectral cues may be more important in speech perception than previously believed.

Cochlear Implants and speech perception

- **However**, cochlear implant users continue to experience difficulty in **noise**, and require much higher signal-to-noise ratios than normal hearing listeners to achieve similar levels of accuracy.

Cochlear implants and noise

- Cochlear implant users may experience substantially greater difficulties in background noise and reverberation. One reason may be that their auditory input has reduced frequency resolution, as a result of the limited number of implant electrodes.

Cochlear implants and noise

- Against a background of speech-shaped noise, normal-hearing listeners presented with speech processed through an implant simulation require more channels to reach the same performance levels (Dorman et al., 1998; Fu et al., 1998).
Cochlear implants and noise

- **More frequency channels** are needed to understand speech in noise. This suggests that reduced frequency selectivity has a greater impact in noise.
- Consistent with this idea, studies have shown that spectral smearing is more harmful to speech recognition in noise than in quiet (ter Keurs et al., 1992; Baer & Moore, 1993).

Effects of background noise

- **Hearing-impaired listeners** often exhibit poorer frequency selectivity than normal-hearing listeners, and they often report difficulty understanding speech in noise.
- They are less able to benefit from spectral and temporal dips in the masker to improve their detection of a target signal, including speech (Festen and Plomp, 1990).

Effects of background noise

- **Cochlear implant** users require higher signal-to-noise ratios than normal listeners to reach a target level of speech recognition.
- Possible reasons:
  1. Limited number of implant electrodes
  2. Limited insertion depth of electrode array
  3. Spectral mismatch from warped frequency-to-electrode allocation
  4. Reduced availability of low-frequency information

Periodicity and noise

- **Hypothesis:**
  - **Periodicity** of speech contributes to robustness
  - **Harmonicity** in the frequency domain
  - Across-frequency grouping of spectral features
  - Unvoiced sounds (e.g., whispered speech) are more susceptible to masking and interference by competing sounds

**Brungart et al. (2001)**

![Graph showing 2-talker correct responses (%)]

Target-to-Masker Ratio (dB)

-12 -9 -6 -3 0 3 6 9 12

2-talker correct responses (%)
Qin and Oxenham (2003)

- Speech recognition performance in noise measured as a function of target-to-masker ratio, processing condition (4-, 8-, or 24-channel CI simulation, or unprocessed) and masker type (SSN, AM noise, same-sex single talker, different-sex single talker).
- Poor performance with implant simulations for normal listeners when speech is presented in fluctuating maskers.

Masking and interference

- **Energetic masking** – reduced audibility of signal components due to overlap in spectral energy within the same auditory filter.
- **Informational masking** – reduced audibility of signal components due to non-energetic factors such as target-masker similarity.
  - Forward vs. backward speech similarity.
  - Familiar vs. foreign language.