DAT330 – Principles of Digital Audio
Cogswell Polytechnical College
Spring 2009

Week 7 – Class Notes

Perceptual Coding

Perceptual coding concerns with the reconstruction of a wave form by favoring the perceived identity, rather than the physical identity (samples). By using psychoacoustic models of the human auditory system, the codec (coder-decoder) identifies imperceptible signal content as bits are allocated and the signal is efficiently coded in the final bitstream. In this way, the data amount needed to represent an audio signal is reduced. However, quantization noise is increased with this process. As a solution to this issue, much of the quantization noise can be shaped and hidden below the signal-depend thresholds of hearing.

PCM, compared to new perceptual coding methods, is considered to be rather inefficient and not suitable for some applications. Perceptual coders open up new possibilities for digital audio and video applications. By understanding how information is perceived by the ear, we can apply this knowledge to create perceptual coding systems that are based on how the human ear perceives sounds.

Psychoacoustics

Psychoacoustics is the study of human auditory perception, ranging from the physiology of the hearing system to the psychological interpretation of aural information. This area of study explains the subjective response to everything we hear by reconciling acoustical stimuli with the physiological and psychological responses evoked by them.

The ear is astonishingly acute in its ability to detect nuances or defects in a signal, but is also rather casual with some of its aspects. The accuracy of a coded signal can be very low, but the degree of accuracy depends on both frequency and time.

Psychoacoustics consider both physical and perceptual measurements. For example, intensity is an objective physical measurement of magnitude. Loudness is the perceptual description of magnitude that is dependent on intensity and frequency. Because loudness cannot be empirically measured, it necessarily must be determined by listener's judgments. Loudness can be expressed in loudness levels called *phons* (the intensity of an equally loud 1 kHz tone, expressed in dB SPL), or it can also be expressed in *sones* (loudness ratios).

The ear's dynamic range is remarkably large; the threshold of loudness has a ratio of 1,000,000,000,000:1 or expressed in decibel from 0 dB (threshold of hearing) to 120 dB (threshold of pain). However, sensitivity to loudness is frequency dependent and are clearly represented by the equal-loudness contours.

Frequency is a literal measurement, while pitch is a perceptual measurement. Pitch is determined by frequency, waveform, and intensity. The ear is most sensitive to frequencies in the range of 1-5 kHz. The ear's response to frequency is logarithmic. For example, the interval of the octave (2:1 ratio). Consider the frequencies interval of 100 and 200 Hz and the interval of 1000 and 2000 Hz. The frequency ratios are the same and are perceived as an octave, but the second octave is much larger than the first. For this reason, musical notation uses a logarithmic measuring scale.

Beat frequencies occur when two nearly equal frequencies are sounded together. The difference in frequency between the tones can be perceived as audible tone or difference tone. Difference tones are easily perceived when the frequencies are high, the tones are fairly loud, and the interval is no larger than a fifth.

The spatial perception of sound sources can be accomplished with several methods: for sounds coming from the sides interaural (ear to ear) intensity differences, time delays, and waveform complexity. For sounds coming from two sources (left-right) located in front of the listener, the perceived sound will be coming from a center space between sources.

The ear perceives only a portion of the information in an audio signal; the perceived portion is the perceptual entropy, estimated to be as low as 1.5 bits/sample. Small entropy signals can be efficiently reduced, while large ones cannot. For this reason, codecs must be able to output variable bit rates so that the bit rate is low for poor information and high for rich information. While the output sampling rate is constant, the entropy of a waveform is not. By using psychoacoustic methods irrelevant portions of a signal can be reduced (data reduction) and the result is inaudible. The codec models the receiver (human ear) in order to identify irrelevant or redundant data in the audio signal.

Physiology of the Human Ear and Critical Bands

The ear is transducer that converts acoustical energy to mechanical energy and ultimately to the electrical impulses sent to the brain, where information contained in sound is perceived. The ear can be divided into three sections: the outer, middle, and inner ear.

The outer ear collects sound and helps assess directionality. Is it comprised of the pinna and the ear canal. The ear canal resonates at around 3-4 kHz providing extra sensitivity crucial fro speech intelligibility.

The middle is is comprised by the ear drum and the ossicles. The ear drum converts acoustical energy into mechanical energy. The ossicles or ear bones (hammer, anvil, stirrup) , the smallest bones in the body, provide an impedance matching mechanism to efficiently convey sounds in the air to the fluid filled inner ear.

The inner ear is formed by the vestibular canals and the cochlea. The vestibular canals are responsible for the motion detection mechanism that provides the sense of balance. The cochlea is mainly comprised of the basilar membrane. This membrane behaves in a analogous manner as a frequency and amplitude analyzer (Fourier analyzer). The vibrations detected are then converted to electrical impulses and sent to the brain as neural information along a bundle of nerve fibers. The brain then decodes the period of the stimulus and point of maximum excitation along the basilar membrane to determine frequency; activity in surrounding local regions is ignored.

The basilar membrane has around 30,000 hair cells arranged in multiple rows, these cell rows are called the Organ of Corti. The cells detect local vibrations of the basilar membrane and convey the information to the brain via electrical impulses. Frequency discrimination dictates that at low frequencies, tones a few Hz apart can be distinguished; at high frequencies, tones must differ by hundreds of Hz. The hair cells responding to the strongest stimulation in their local region is called the critical band (CB). Critical bandwidth is not fixed, it increases with increasing frequency. Three-fourths of the critical bands are located below 5 kHz, which means that the ear receives more information from low frequencies than for high frequencies. Critical bandwidth can be approximated by using:

$$\mathrm{CB\ (Hz)} = 25 + 75[1 + 1.4(f/1000)^2]^{0.69}$$

Physiologically, each critical band occupies a length of about 1.3mm, with 1300 primary hair cells. The critical band concept describes the filtering process occurring in the ear, which is basically that of a spectrum analyzer with response patterns of overlapping bandpass filters with variable center frequencies.

Critical bands have been also used to explain consonance and dissonance. Intervals with a difference greater than a critical band are considered consonant, while intervals of less than a critical bandwidth are considered dissonant. Dissonance tends to increase at low frequencies. Finally, critical bands play a role in the perception of pitch, loudness, phase, speech intelligibility, and other perceptual matters.

The Bark is a unit of perceived frequency, specifically it measures the critical-band rate. The Bark scale relates absolute frequency (Hz) to perceptually measured frequencies such as pitch or critical bands (Bark). Converting frequency to Bark is done with:

$$z(f) = 13\arctan(0.00076f) + 3.5\arctan[(f/7500)^2] \text{ Bark}$$

By using the Bark scale, the physical spectrum can be converted to a psychological spectrum along the basilar membrane. The shape of a masking curve varies with sound level, becoming more asymmetric with louder levels.

The shape of an auditory filter can also be described in term of an equivalent rectangular bandwidth (ERB) scale. The ERB can be modeled by:

$$\text{ERB} = 24.7[4.37(f/1000) + 1] \text{ Hz}$$

The pitch place theory explains the action of the basilar membrane in terms of frequency-to-place transformation. The vibration of the sound wave will create a peak in a particular place along the basilar membrane. High frequencies cause peaks near the middle ear, while low frequencies cause peaks at the far en of the membrane. Hair cells located near the place of strongest stimulation will convey that frequency in a critical band.

Threshold of Hearing and Masking

The threshold of hearing curve describes the minimum level at which the ear can detect a tone at a given frequency. In other words, it describes the energy needed in a pure tone to be barely audible in a noiseless environment. In general, two tones of equal power and different frequency will not sound equally loud. Sensitivity decreases at high and low frequencies.

A perceptual codec compares the input signal to the minimum threshold and discards signals falling bellow the threshold. Likewise, the codec can place quantization noise far below the threshold. The threshold is absolute, but because sound can be played back at different loudness levels codecs have to compare the lowest output level of the decoder with the the threshold curve.

Amplitude masking occurs when a loud tone completely obscures softer tones. Masking shifts the threshold curve upward in a frequency region above the curve. The masking threshold, thus describes the level in which a tone is barely audible. The strong sound is called the masker, while the softer sound is called the maskee. Masking theory suggests that the softer tone is detectable only when its energy is equal to the energy of the masking signal in the critical band.

The mechanics of the basilar membrane explain the phenomenon of masking. A loud response at one place of the membrane will mask a softer response in the critical band around it. Unless the activity from another tone rises above the masking threshold, it will be swamped by the masker.

Temporal Masking

Occurs when the tones are sounded sequentially. Two types of temporal masking exist: backward masking, signal precedes the masker; and forward masking, the masker is presented first and the signal follows it. Temporal masking increases with the reduction of time differences.

In frequency domain coding, temporal masking is important because the codecs have a limited resolution because they operate on blocks of samples, thus spreading the quantization error over time. Temporal masking can reduce the audibility of artifacts caused by transient signals. Filter banks in addition to masking can for a contour mapped in the time-frequency domain, such that sounds falling under this contour will be masked. Perceptual codecs must then identify this contour for changing signal conditions and code the signal appropriately.

Rationale for Perceptual Coding

Perceptual codecs maintain a constant sampling frequency, but selectively decrease the word length. The word length reduction is done dynamically based on signal conditions. Masking and other factors are considered so that the resulting increase in quantization noise is rendered as inaudible as possible. The level of quantization error, resulting from truncating the word length, can be allowed to rise as long as it can be masked by the audio signal.

Perceptual codecs analyze the frequency and amplitude content of the input signal and compare it to a model of human auditory perception. Using this model, the codec is able to remove irrelevant and redundant data from the audio signal. Although this method is lossy, the listener will not perceive any degradation in the decoded signal. Considerable data reduction can be achieved, up to 83%, depending on the codec used. Well-designed codecs can rival the sound quality of conventional recording techniques.

The efficiency of a perceptual codec rises from the adaptive quantization used. PCM signals are assigned equal sized word lengths. Perceptual codecs assign bits according to audibility. More bits are given to a prominent tone, while fewer bits are used to code softer tones and none to inaudible tones. A codec's reduction ratio (coding gain) is the ratio of input bit rate to output bit rate. Reduction ratios of 4:1, 6:1, or 12:1 are common.

The heart of the perceptual codec is the bit allocating algorithm. Its function is to determine how best to distribute the bits across the signal's spectrum and requantize samples to minimize audibility of quantization noise while meeting the overall bit budget for that block.

Two bit allocation strategies can be used in perceptual codecs: forward adaptive allocation and backward adaptive allocation.

Forward adaptive allocation performs the bit allocation in the encoder and contains the encoding information in the bitstream. Its allocation can be very accurate, depending on the encoder, and has the advantage that the psychoacoustic model is located in the encoder. The decoder does not need the model since the encoded data is used to reconstruct the signal. A disadvantage is that a portion of the available bit rate is needed to convey the allocation information of the decoder.

Backward adaptive allocation derives the bit allocation information from the coded audio data itself without explicit information from the encoder. The bit rate is not partly consumed by allocation information. However, accuracy might be limited because the bit allocation in the decoder is calculated from limited information.

Perceptual coding is generally tolerant of errors. The error is limited to a narrow band corresponding to the bandwidth of the coded critical band, thus limiting loudness. Error might be perceived as a burst of low-level noise. Perceptual coding also permits targeted error correction. For example, more vulnerable sounds (soft sounds) are given greater protection than less vulnerable sounds (loud sounds). Perceptually coded data requires appropriate error correction for storage or transmission.