A Comprehensive Overview into the Psychoacoustic Phenomenon of Auditory Masking

Adar Guy
University of Victoria

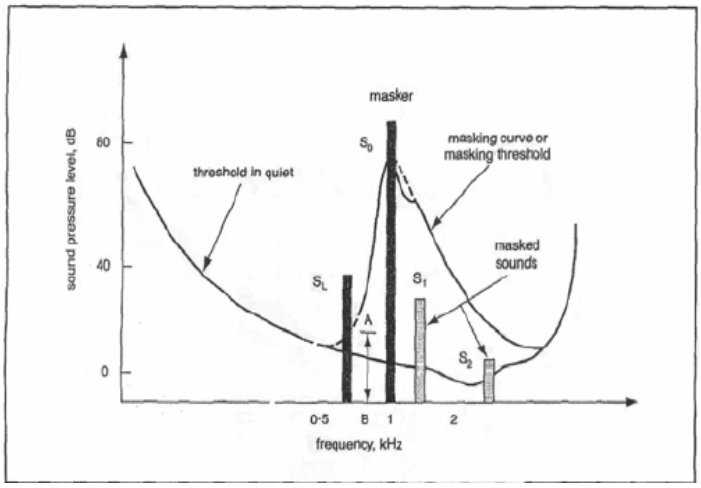
Introduction

It is quite common to hear more than one sound at a given time. The experience of listening to someone speak, or hearing a piece of music, can often be made more challenging by the presence of interfering sound, be it surrounding conversations, a car passing by, or any other noise that may occur in a given space. In some situations, the background noise is so loud that it interferes with a listener's ability to hear their target sound. This phenomenon is known as masking, and it occurs in every sensory system. Auditory masking occurs when the perception of one sound is affected by the presence of another sound. This psychoacoustic effect exists for all humans (to different degrees) and can often be used in a number of ways to the researcher or engineer's advantage. The aim of this paper is to present a comprehensive overview into the psychoacoustic phenomenon of auditory masking.

Definition

Masking is the process by which the detection threshold of a signal is increased by the presence of a stronger signal (masker). A 'not-masked' threshold is the quietest level of the signal which can be perceived when isolated. A masked threshold is the quietest level of a signal that can be perceived when accompanied by noise. The 'amount of masking' is the difference between the two, and is defined as the increase (in dB) in the detection threshold of a signal due to the presence of a masker (a non-masked threshold of 20dB and masked threshold of 15 dB means an amount of masking of 5 dB) [Patel]. Masking is often used to investigate the auditory system's ability to separate the components of a complex sound. Two signals with a large enough difference in frequency will produce two

separate pitches when played at the same time. This is known as frequency selectivity and it occurs because of filtering in the cochlea. However, the limits of frequency selectivity in the basilar membrane can be determined using masking if the auditory system is unable to distinguish between the two frequencies. An experiment can be implemented to determine the amplitude thresholds (the conditions which are necessary for one sound to mask a previously heard sound), investigating the frequency selectivity of the auditory system, and determining amplitude thresholds for a given signal [Bregman]. This is crucial to the perceived output of the superimposed signal and varies depending on the masking signal used. Masking can occur both in the frequency domain (leading to frequency analysis model) and in the time domain (pre/post masking).

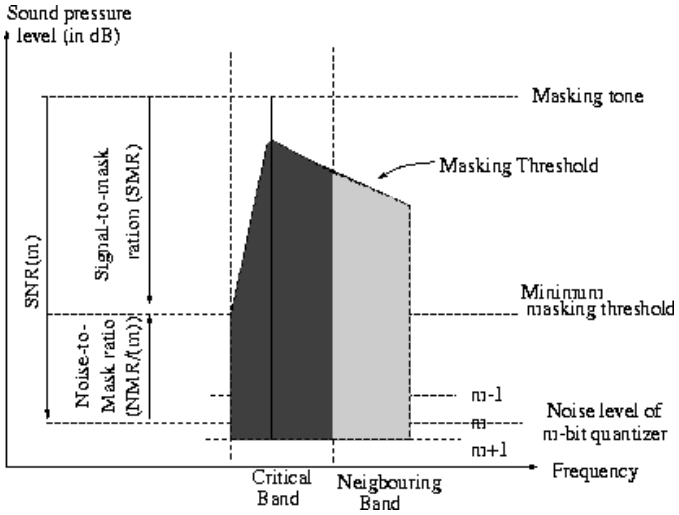


Some general rules for masking must also be established. Lower tones will effectively mask higher tones (but the opposite is not true). The greater energy that the masking signal has, the wider its influence is, which means that it has a broader range of frequencies that it can mask (but as a consequence, if two tones are widely separated in frequency then little masking will occur). The greater the difference between the masking signal energy and the masked signal energy, the larger the stimulation will be and it will take longer for the energy to dissipate (persistence) [Moore]. Lastly, masking thresholds increase with increased bandwidth of the masker, simply because more energy is available for masking.

Simultaneous Masking

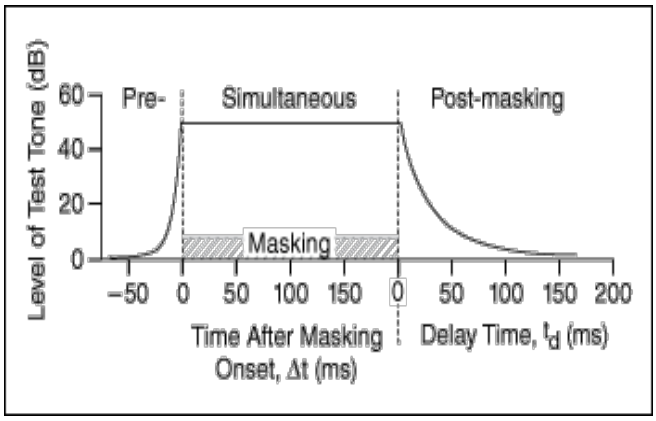
In the frequency domain, masking occurs simultaneously. This is an intuitive phenomenon where a low level signal (small-band noise), can be made inaudible by a simultaneously occurring stronger signal. The masking threshold depends on the frequency of the masker, sound pressure level, and the characteristics of both signals. The slope of the masking threshold is steeper towards lower frequencies meaning that higher frequencies are masked more easily [Brandenburg].

Without a masker, a signal is inaudible if its SPL is below the threshold of quiet, which depends on frequency and covers a dynamic range of more than 60 dB – but this is only masking by one masking signal. If the source signal consists of many simultaneous maskers, a global masking threshold is determined based on a high resolution short-term amplitude spectrum of the signal (sufficient for critical band analysis). It maps the threshold of just noticeable distortions as a function of frequency. This is accomplished by first calculating individual masking thresholds (based on signal level, type of masker, and frequency range), and then summing them together along with the threshold in quiet (used as a buffer to ensure that resulting threshold is above quiet threshold) [Bregman]. Lastly, the signal-to-mask ratio of the global threshold must be calculated by determining the ratio of the maximum signal power and global masking threshold. The diagram on the right shows this ratio. When calculating the sum of individual masking thresholds, the effects of masking reaching over critical band bounds must be included in the calculation.



Temporal Masking

In addition to simultaneous masking there also exists two other time-domain phenomena that play an important role in human auditory perception; pre-masking and post-masking. These temporal masking effects occur before and after a masking signal has been switched on and off. The first time domain masking effect is called post-masking or forward masking. Post-masking is associated with 4 neurophysiological properties that exist in the auditory system. The first is central inhibition and it is the result of the interaction of processes at the level of the neuron, and nerve center. In this case it is a reorganization of the neural structure that exists in a state of excitation or

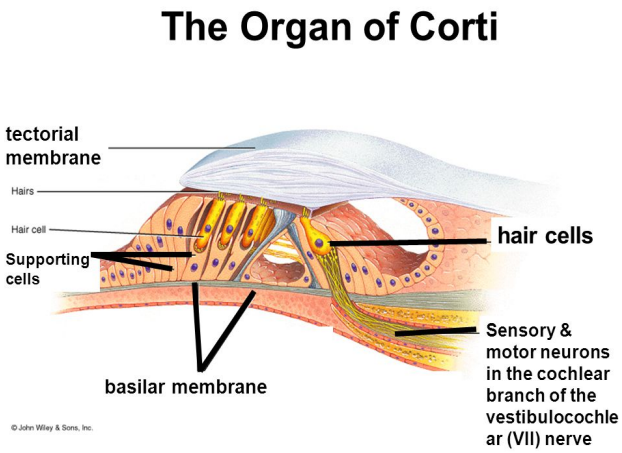


delayed inhibition. Next, is the continual response of the basilar membrane. It does not end immediately after the offset of the stimulus, instead persisting over a period of time. Third is the neural response which acts similarly to the basilar membrane as it also continues for a short period. Lastly, nerve fiber adaptation causes a signal to invoke a higher nerve response if the signal is more isolated in time. This means that a signal that is close to another (masking) signal will cause a smaller neural response than that of a more separated (isolated) signal to the masker [Moore]. Pre-masking or backward masking is the other time domain phenomenon of which little is known. The duration when pre-masking applies is significantly less than one tenth that of post-masking, which is in the order of 20 to 200 msec. It has some very important roles in production for hiding unwanted artifacts like pre-echoes. It along with post masking are used to develop ISO/MPEG audio coding algorithms.

Anatomy

Masking takes place in the first stages of electro-mechanical processing of the cochlea. The sound waves of multiple sources become superimposed, interleaving and overlapping in time and frequency which, in strict mathematical terms, cannot be 'unmixed'. The fluctuation in air pressure produced by the combined signal causes a vibration in the eardrums. The auditory system must decompose the pattern of vibrations to gather information about the individual sound sources, namely identity and location. The components are then coded on the auditory nerve which transmits sound information to the brain [Pletsch]. The ability to discriminate and separate the combined signal into its separate sources is accomplished with the process of auditory scene analysis. A complex sound is separated into different frequency components causing peaks in the vibrational pattern on a specific point of the basilar membrane. Accomplishing this task can be difficult because the ear only has access to the summed pressure wave coming from separate sound sources, but using heuristic processes, the brain constructs separate spatial and identificational descriptions of each sound source. For example, one such heuristic analysis exists in the time domain and determines source separations based on temporal occurrence. If one subset of frequency components begins together at exactly the same time, whereas another subset that also begin together, but at a different time from the first subset, then each subset will be considered to originate from separate sources. Other heuristic analyses may be based on the different regularities in how a sound (signal) is produced. Collecting these components can determine the perceived pitch, timbre, loudness, and spatial position of the sound [Pletsch]. This ability of the auditory system is perhaps one of the most complex and impressive achievements of the brain.

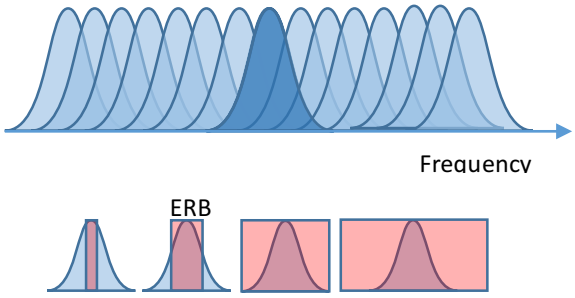
The decomposition of vibrational patterns occurs in the cochlea, where different frequencies within a complex signal stimulate different places along the basilar membrane (a portion of the ear that runs along the cochlea). The spatial decomposition or frequency-to-place mapping is known as tonotopy and establishes the fundamental organizational principle of auditory coding. Studies in physiological and perceptual experiments typically involve embedding a tonal signal in a noise masker [Bregman]. This sound level of the noise is the 'masked threshold'. Response to noise stimulus of the mammalian basilar membrane shows that the frequency-specific filtering of noise and tones as well as other non-linear neural phenomena produced by the interaction of tones and noise, originate in the basilar membrane's mechanical response [Pletsch].



Linear aspects of auditory processing, such as the constant signal-to-noise ratio at thresholds over a large dynamic range, also resemble early correlations of the physiological perceptual invariance experienced when listening to the same sound over a wide range of stimulus levels. It exists throughout the lower sections of the auditory system like the auditory nerve, up to and including the primary auditory cortex. Masking occurs when the tonotopic representation of one sound interferes with that of another. The experiments conducted on the physiological origins of masking generally observe that both in the neural and psychophysical states, there is an established level of mechanical transduction in the cochlea [Patel]. Two very important and controversial questions regarding this investigation still remain. The first is to what extent the cochlear responses of complex sounds are altered by efferent control (activated by prior sounds, or 'top-down' effects like attention). The second more controversial question is how similar the human cochlear tuning system is to that of other mammals that are studied for their similarity to humans.

Investigation into the physiological origins of masking can also provide beneficial information to use in testing and validating non-linear computational models of the peripheral auditory system [Patel]. The results can be used as front ends for devices like automatic speech recognizers or low bit-rate audio codecs (such as those used for MP3 audio compression). Models for perceptually transparent compression and audio coding are based on a masking function which distributes the quantization noise in the least

sensitive regions of the spectrum (typically high frequency regions) to minimize its perceptual effect. As previously stated, auditory perception is based on critical band analysis, which is a frequency to location mapping occurring on the basilar membrane. Instead of being represented on a linear frequency scale, the power spectra are represented using limited frequency bands called critical bands. In 1940 Fletcher stated that the auditory system behaved like a bank of overlapping band pass filters called "auditory filters", with bandwidths on a perceptual frequency scale in the order of 100 Hz for signals below 500 Hz and up to 5 kHz for high frequencies (up to 24 kHz 26 critical bands must be considered). He stated that masking thresholds occur when the acoustic power of the signal at the filter output is proportional to the acoustic power of the masker at the filter output [Moore].



The shape of the filter can be determined by looking at the shapes of psychoacoustic tuning curves. The critical bands define the frequency analysis model that the masking function is characterized by. The amount of masking increases with increasing masking energy that gets through the filter up to the bandwidth of the filter and then any increase in noise energy (bandwidth) does not effect the masking energy. The most efficient mask-filter ratio called an equivalent rectangular bandwidth (ERB) which has an equal height and total area to the filter (same energy) but different shape [Moore].

Masking in Application

In application, masking can become a very useful tool. The logical basis for this is rooted in the purpose of music and sound production. It is ubiquitous that music and other sound production is designed primarily (if not purely) for use by humans. Thus, any models which can be derived from psychoacoustic or neural phenomena would most accurately resemble our perception and would produce the most efficient systems that represent our needs as listeners. Each type of masking is used in various applications appropriate to the needs of the application and the capabilities of that masking phenomenon.

As previously mentioned, pre-masking or backward masking is often used to hide unwanted artifacts such as pre-echo. Pre-echo is a transform based compression artifact in digital audio that commonly occurs in impulsive, transient sounds like percussive instruments. An echo is heard preceding the transient but is then masked by the transient so no resulting post-echo is perceived. It is caused by a spreading of quantization noise

over the codec's transform window – quantized transform coefficients produce noise in all instances of the time domain and so a quiet enough signal block with a belated transient like a cymbal will cause audible noise before the transient [Patel]. If the transform blocks however are made short enough, this effect can be made inaudible with pre-masking.

Filters are used in this case to produce response only occurring after the transient (instead of linear phase filters). Although the filters introduce their own set of artifacts like temporal smearing and phase distortion, these artifacts are less audible as they are dealt with by strong post-masking. In transform domain lossy audio codecs, often relying on MDCT, avoiding pre-echo can have considerable design difficulties that are also encountered in digital room correction algorithms and general filter design for spectral equalization, subtraction and more.

Masking energy that lies outside the spectral region of a targeted signal can improve detection and identification of that signal. This phenomenon is known as comodulation masking release and it is essentially a decrease in the masked threshold of a signal when the masking signal is amplitude modulated [Piechowiak]. This happens despite the fact that the masking bandwidth is increased during amplitude modulation which as previously stated, would normally cause an increase in masked thresholds. This interesting feature of CMR has been used in the advancement of cochlear implant technology. Speech identification and discrimination is enhanced when applying a comodulated flanking band, but research has proven that CMR appears to be reduced or absent in persons suffering from cochlear hearing loss – this discrimination inability causes a difficulty in understanding speech in noisy backgrounds [Piechowiak].

Perception based coders use encoding processes that are controlled by a ratio of a global signal-to-mask ratio over a frequency curve. In order to produce a transparent coding scheme, an appropriate bit rate for masking of all distortions must be available, otherwise the global masking threshold acts as a spectral error weighting function that will shape the error spectrum. Perceptual coding designs attempt to reach the limits of just noticeable distortions using models which map neural impulses and psychoacoustic phenomena. Perceptually coded material may also sound better if dynamic bit allocation were used [Brandenburg].

Lossy audio data compression like MPEG for example, removes sounds that are already masked anyway. MPEG applies a filter bank to the input to break it up into its frequency components as a psychoacoustic model is implemented to the data for bit allocation. The number of bits allocated are used to quantize information from the filter

bank which provides the compression [Ambikairajah]. MPEG defines 3 layers for audio where the basic model is of the same complexity as that of the codec, which increases with each layer. In the first layer there is only 1 frame in a DCT (discrete cosine transform) filter. Equal frequency is spread per band and the psychoacoustic model only uses simultaneous frequency masking. The second layer consists of 3 frames in the filter (before, current, next) developed by the psychoacoustic model which uses both pre and post masking. The third layer is MP3 and it uses a good critical band filter which accounts non-equal frequencies and takes into account stereo redundancy. It uses a Huffman coder (prefix frequency based mapping algorithm) and a psychoacoustic model that brings together both temporal and frequency masking [Ambikairajah].

It is likely that all future coding schemes will make use of psychoacoustic models, similar to the advancement and use of AI in other technologies – they both attempt to model natural actions or occurrences of the human mechanical system. The motivation in continuing to advance perceptual coding may have to do with file compression for transferring over a network in less time, seeing as music is being bought and sold (transferred) more readily online. As more research on masking is conducted, it will continue to be exploited for perceptual coders and other applications. It is truly an interesting phenomenon that is fundamental to defining us as human beings.

Work Cited

Ambikairajah, E., A G. Davis, and W.T K. Wong. "Auditory Masking and MPEG-1 Audio Compression.". Accessed 9 Dec. 2016. personal.ik.itba.edu.ar/~nbaum/Clase%20Audio%20Compression/Auditory%20masking%20and%20MPEG-1%20audio%20compression.pdf

Brandenburg, Karlheinz. "Introduction to Perceptual Coding*." *Collected Papers on Digital Audio Bit-Rate Reduction*, 1996. *AES*. Accessed 9 Dec. 2016.

Bregman, Albert S. "Auditory Scene Analysis." *International Encyclopedia of the Social and Behavioral Sciences*, Accessed 9 Dec. 2016. webpages.mcgill.ca/staff/group2/abregm1/web/pdf/2004_%20Encyclopedia-Soc-Behav-Sci.pdf

Bregman, Albert S. "When Will We Hear Separate Events in a Sequence of Sounds?" *AES 103rd Convention*, 1997. *AES*. Accessed 9 Dec. 2016.

Moore, Brian C. *Hearing: Handbook of Perception and Cognition*. Second ed., London, Academic Press Limited, 1995, pp. 161-203. MIT, University of Cambridge. Accessed 9 Dec. 2016.

Piechowiak, Tobias, Stephen D. Ewert, and Torsten Dau. "Modelling comodulation masking release using an equalization-cancellation mechanism." *Acoustical Society of America*, 2007. Accessed 9 Dec. 2016. orbit.dtu.dk/files/4300590/piechowiak_et_al.pdf

Patel, Aniruddh D. *Music, Language, and the Brain*. Second ed., New York, Oxford University Press, 2008.

Pletsch, Brandon, Narrator. *Auditory Transduction*. Brandon Pletsch, YouTube, 2009. Accessed 9 Dec. 2016. https://www.youtube.com/watch?v=46aNGGNPm7s