ΨΗΦΙΑΚΗ ΕΠΕΞΕΡΓΑΣΙΑ ΣΗΜΑΤΩΝ 2022

Σχολή ΗΜ&ΜΥ, ΕΜΠ

Audio Coding/Compression Tutorial

DSP22-LAB2

Overview

- Digital sound
- Audio coding
- MPEG basics
- Basic Encoding Architecture (with Lab instructions)
  - Psychoacoustic Model
  - Filterbank
  - Quantization

Digital audio encoding

- Το CD και ο ψηφιακός ήχος έχουν ουσιαστικά αντικαταστήσει τον αναλογικό ήχο
- CD: Εξαιρετική υψηλή πιστότητα (high fidelity), δυναμική εμβέλεια καθώς και ευρωστία, σε βάρος όμως του μεγάλου ρυθμού μετάδοσης/μεταφοράς των δεδομένων
- Οι διάφορες εφαρμογές ήχου (π.χ. ασύρματη μεταφορά ή πολυμεσικά συστήματα) απαιτούν: μειωμένο bandwidth, μικρή χωρητικότητα αποθήκευσης και χαμηλό κόστος

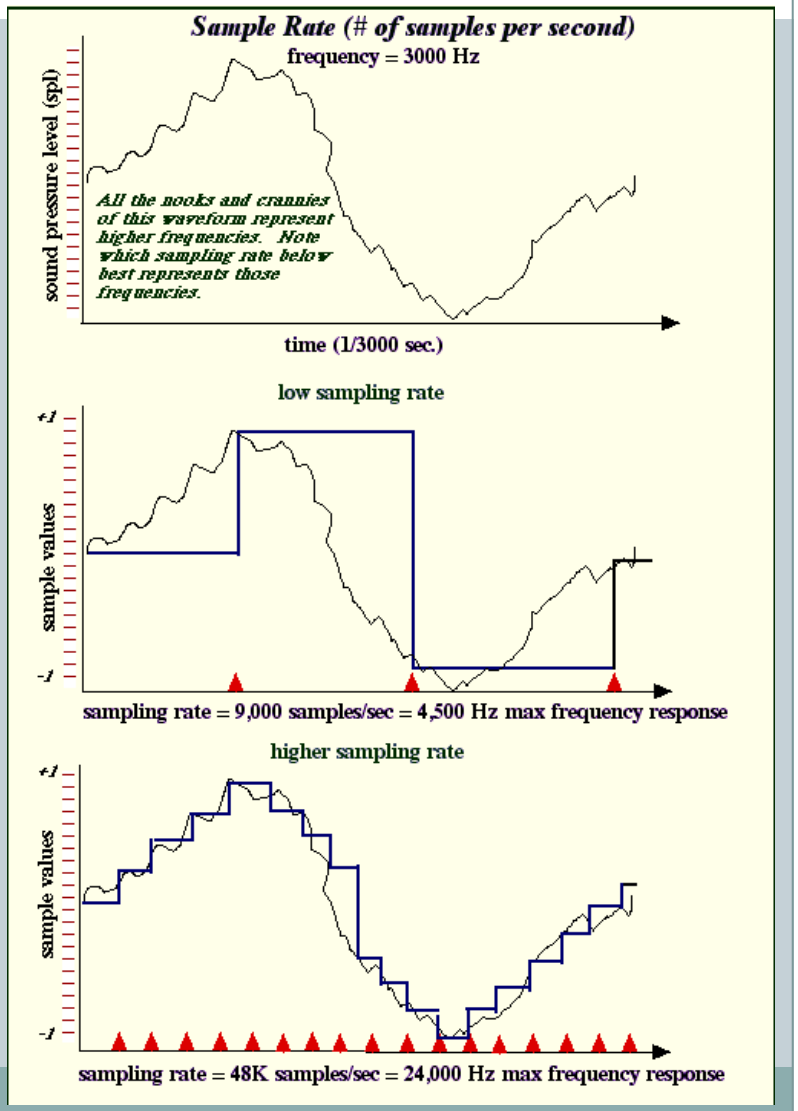
➢ **Ζητούμενο**: η μετάδοση ήχου υψηλής ποιότητας με χαμηλά bit-rates

➢ **Ανάγκη**: αλγόριθμοι για την κωδικοποίηση του ψηφιακού ήχου με αντιληπτικά διαφανή τρόπο σε ποιότητα που να φτάνει αυτή του CD

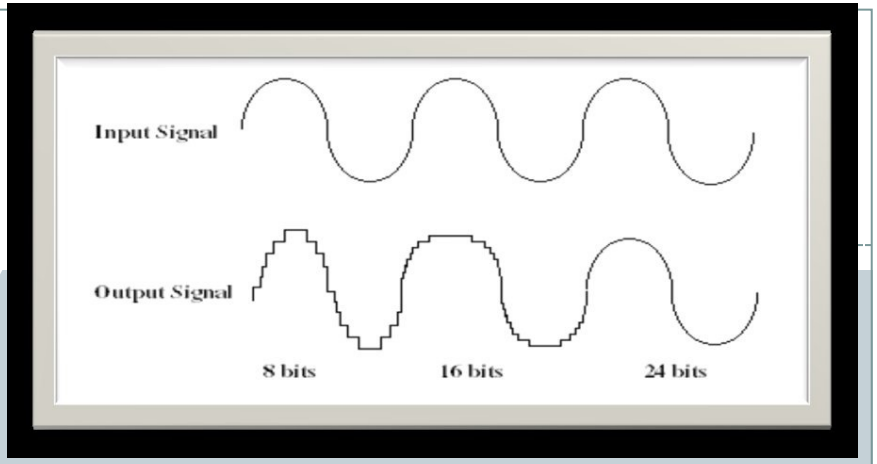
Sampling rate/frequency ($f_s = 1/T_s$)

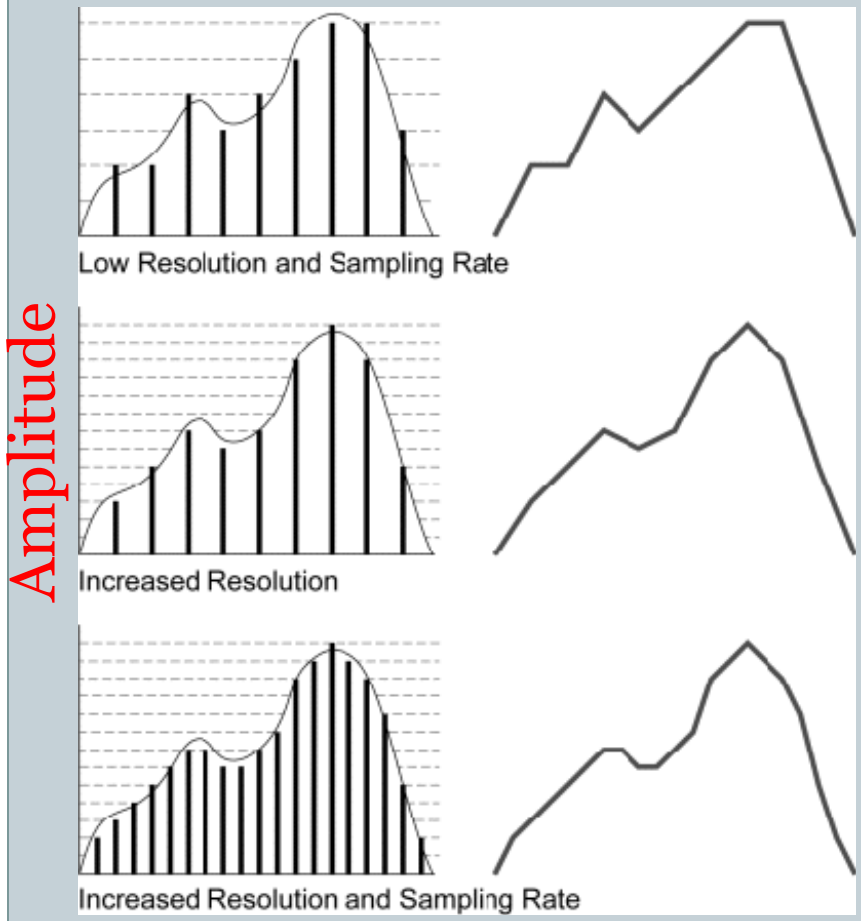
- **Sampling Rate ή SR:** ο αριθμός των δειγμάτων ή μετρήσεων ανά δευτερόλεπτο.
  - Στο CD η συχνότητα αυτή είναι 44100 Hz. Το δείγμα δηλαδή χωρίζεται 44100 το δευτερόλεπτο (οτιδήποτε μεσολαβεί μεταξύ δύο διαδοχικών μετρήσεων, αγνοείται και δεν καταγράφεται).
- **Νόμος του Nyquist:** το SR πρέπει να είναι τουλάχιστον διπλάσιο της μέγιστης συχνότητας.
- Μέγιστη συχνότητα του ακουστικού μας φάσματος είναι ca. 20KHz, άρα το SR πρέπει να είναι τουλάχιστον 40.000Hz.

- **Μεγαλύτερος ρυθμός δειγματοληψίας:** μετρήσεις σε μικρότερα διαστήματα και, καθώς ο ήχος είναι συνάρτηση της συχνότητας, μπορεί να καταγράφει και υψηλότερες συχνότητες.
- **Χαμηλή συχνότητα δειγματοληψίας:** η παραγόμενη καμπύλη δεν ακολουθεί πιστά την αρχική κυματομορφή, δημιουργείται σφάλμα και παραμόρφωση αλλά και η αρμονική παραμόρφωση **aliasing.**



Bit Sample Resolution





- Μέτρηση του ύψους της κυματομορφής = ένταση (amplitude) κάθε στιγμής.
- Μεταξύ μέγιστης και ελάχιστης τιμής υπάρχουν άπειρες διαβαθμίσεις, που ο υπολογιστής δεν μπορεί να αναγνωρίσει.
- => Άρα χωρίζει την περιοχή σε ίσα μέρη και καταγράφει πεπερασμένο πλήθος μετρήσεων έντασης.
- 1 byte (8 bits): 256 δυνατές τιμές **(2^8)**.
- Θέλουμε αρκετά bits ώστε το σφάλμα κβάντισης να είναι μικρό και άρα η κυματομορφή που παράγεται να μην απέχει αρκετά από την αρχική.
- Στο CD για κάθε μέτρηση αποθηκεύονται 2 byte (16bit) και μπορούν να εκφράσουν 256*256=65.536 διαφορετικές τιμές έντασης.

High quality audio (CD)

44.1 kHz sampling rate with 16 bit sample resolution (256x256 = 65636 values of amplitude)

PCM (Pulse Code Modulation) Audio

This means:

- bit rate of 706.5 kilobits/sec (kbps) per channel (monaural)
- bit rate of 1.41 megabits/sec (mbps) for stereo

  (2 * 44.1 kHz * 16 bits= 1.41 Mb/s)

+Overhead (synchronization, error correction, etc.)

CD Audio=4.32 Mb/s

File size = sampling frequency (Hz) * bit rate (bits) * duration (sec) (*2 if stereo)

1min of CD quality audio = 10MB

Coding at a glance

- Ζητούμενο: αρχεία CD τα οποία και θα κωδικοποιηθούν με τέτοιο τρόπο:
  - Έτσι ώστε να συμπυκνωθεί όλη η πληροφορία και να καταλαμβάνουν λιγότερο αποθηκευτικό χώρο στον υπολογιστή.
- Οι κωδικοποιητές γενικά αυτό που κάνουν είναι να αφαιρούν τμήματα του ήχου (μουσικής) τα οποία δεν ακούμε.
  - Η διαδικασία αυτή και η αφαίρεση κάποιων ήχων δεν επηρεάζει καθόλου το πώς τελικά ακούγεται ένα μουσικό κομμάτι.
- Οι αλγόριθμοι κωδικοποίησης δεν ακολουθούν κάποιο συγκεκριμένο πρότυπο και γι' αυτό κάποιοι από αυτούς παράγουν μικρότερα αρχεία από άλλα.
- Bit rate: Ο αριθμός των bits είναι σημαντικός για την ποιότητα του παραγόμενου αρχείου ήχου (typical minimum bit rate is 128kbit/sec).

What is Audio Coding (compression)?

- Χρησιμοποιείται για να έχουμε συμπαγείς ψηφιακές αναπαραστάσεις ακουστικών σημάτων υψηλής πιστότητας με σκοπό την **αποτελεσματική μετάδοση ή αποθήκευση.**

- Έχουμε **διαφανή αναπαραγωγή του σήματος**: δηλ. η διαδικασία μείωσης του ρυθμού μετάδοσης του ψηφιακού σήματος γίνεται με ελάχιστη ή/και καθόλου απώλεια στην ποιότητα ήχου (*transparent signal reproduction*).

- Η είσοδος στον κωδικοποιητή είναι ένα ψηφιακό σήμα.
- Η έξοδος του κωδικοποιητή είναι και πάλι ένα ψηφιακό σήμα αλλά με μικρότερο σε μέγεθος και bit rate.
- Ο αποκωδικοποιητής αντιστρέφει τη διαδικασία και μας δίνει προσεγγιστικά το αρχικό ψηφιακό σήμα.

Historical Coder "Divisions":

Lossless Coders vs. Lossy Coders

Αναστρέψιμο αποτέλεσμα, τέλεια αναπαράσταση του αρχικού σήματος, i.e., the original signal can be reconstructed bit for bit.

Προσεγγιστική αναπαράσταση του σήματος εισόδου, το λάθος ή αυτό που χάνεται κατά την επεξεργασία εξαρτάται από τη μέθοδο που χρησιμοποιείται.

Numerical Coders vs. Source Coders vs. Perceptual Coders

Almost always a lossless
Uses abstract numerical methods to remove redundancies.

New Lossy Numerical coders can provide fine-grain bit rate scalability.

Examples: Huffman Coding, Arithmetic Coding, Ziv-Lempel (LZW) Coding

Lossless or lossy.

a) Remove redundant components (exploit correlations between its samples / information that can be reconstructed)
b) Remove components irrelevant to the ear

Examples: LPC, Sub-band, Transform, Vector Quantization

Incorporating psychoacoustic knowledge of the auditory system.

Irrelevancy: Remove parts of the signal that the human cannot perceive.

In practice, most perceptual coders attempt to remove both irrelevancy and redundancy.

➔ provide the lowest bit rate possible in order to give audible quality.

MPEG Basics

- What does MPEG-1 Audio provide?

  Βασίζεται στα ελαττώματα του αυτιού (lossy audio compression) και σε μετρικές όπως η **υποκειμενικά μη-διακριτή διαφορά μεταξύ των ήχων.**

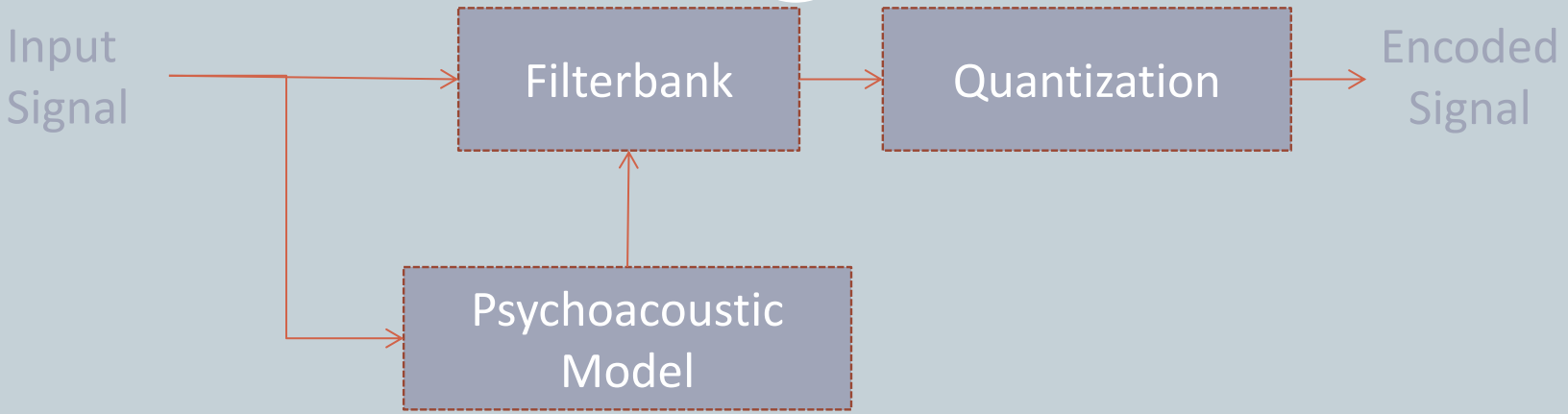
- Μπορεί να συμπιέσει το σήμα κατά έναν παράγοντα 6 και να διατηρήσει την ποιότητα.

- Αποτελεί στην πραγματικότητα μέρος ενός προτύπου (standard) που εμπεριέχει πληροφορίες για τον συγχρονισμό του ήχου, της εικόνας και της audio-visual πληροφορίας.

MPEG-I Audio Characteristics

- PCM sampling rate of 32, 44.1, or 48 kHz
- Four channel modes:
  - Monophonic and Dual-monophonic
  - Stereo and Joint-stereo
- Three modes (layers in MPEG-I):
  - Layer I: **Computationally cheapest**, bit rates > 128kbps
  - Layer II: Bit rate ~ 128 kbps
  - Layer III: Most complicated encoding/decoding, bit rates ~ 64kbps, originally intended for streaming audio, with best audio quality!!

Encoding Architecture



1. Filterbank: μετατροπή σου σήματος εισόδου στο πεδίο της συχνότητας χρησιμοποιώντας 32 φίλτρα.
2. Psychoacoustic Model: υπολογισμός και εύρεση την μη χρήσιμης πληροφορίας του σήματος, πληροφορίας που δεν ακούμε μέσω του ψυχοακουστικού μοντέλου (masking).
   - Bit Allocator (quantization and coding):
     - Εάν η ισχύς σε μια συχνοτική ζώνη είναι κάτω από το κατώφλι κάλυψης, δεν κωδικοποιείται.
     - Ειδάλλως αποφασίζουμε τον αριθμό των bits που χρειάζονται για την κωδικοποίηση έτσι ώστε ο θόρυβος που θα εισαχθεί από τον κβαντισμό να είναι μικρός και να μην ακούγεται.
     - Format bitstream: αποτελείται από τα κωδικοποιημένα δεδομένα και ορισμένες έξτρα πληροφορίες.

Psychoacoustic Model

- Βασίζεται σε θεωρίες σχετικές με την αντίληψη (ο άνθρωπος δεν ακούει όλες τις συχνότητες):
  - Limitations of human auditory system
  - **Don't code if the ear cannot hear it!**
  - Χρησιμοποιούνται μοντέλα για την ανθρώπινη ακοή για να αφαιρεθούν δεδομένα που δεν μπορούμε να ακούσουμε
- Το εύρος συχνοτήτων που ακούμε είναι περίπου **20 Hz έως 20 kHz**
  - με μεγαλύτερη ευαισθησία στο εύρος των **2 έως 4 KHz**.
- Dynamic range: από το πιο ήσυχο έως πιο δυνατό **περίπου 96 dB**
- Το εύρος φωνής είναι **περίπου 500 Hz έως 2 kHz**
- Χαμηλές συχνότητες -> φωνήεντα, μπάσα; Υψηλά -> σύμφωνα

Psychoacoustic Model

- **Two key properties:**
  - **Threshold of Hearing**

    Describes the notion of “quietness”
  - **Auditory Masking :** describes the situation where a weaker but clearly audible signal (maskee) becomes inaudible in the presence of a louder signal (masker)
    - Frequency Masking

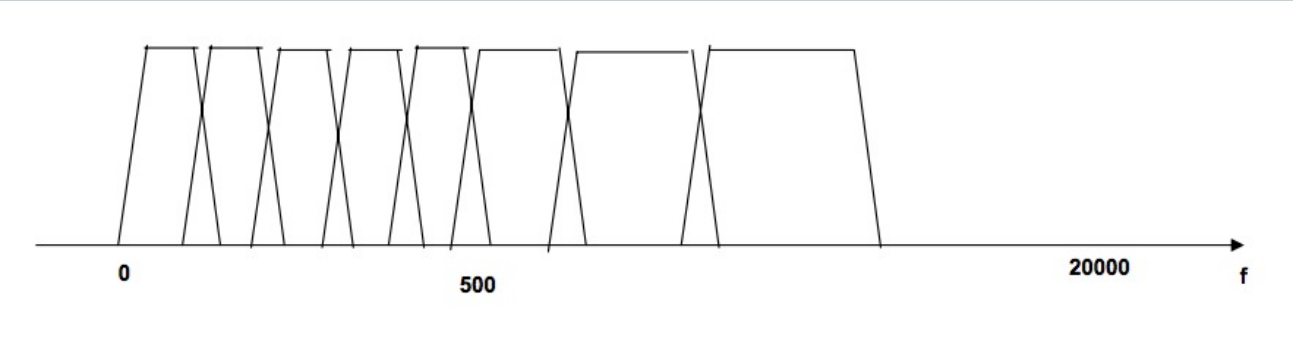
      A component (at a particular frequency) masks components at neighboring frequencies. Such masking may be partial.
    - Temporal Masking

      When two tones (samples) are played close together in time, one can mask the other.

Ear as a Filterbank

- Το σύστημα ακοής του ανθρώπου μπορεί να μοντελοποιηθεί **ως συστοιχία ζωνοπερατών φίλτρων** (25 επικαλυπτόμενα φίλτρα/ζώνες - critical bands) από 0 έως 20 KHz.
- Το αυτί **δεν μπορεί να διακρίνει ήχους που εμφανίζονται ταυτόχρονα** μέσα στην ίδια κρίσιμη συχνοτική ζώνη.
- Κάθε ζώνη (critical band) έχει εύρος ζώνης ca. 100 Hz για ήχους κάτω των 500 Hz, ενώ αυξάνεται γραμμικά πάνω από τα 500 Hz έως και τα 5000 Hz.
- Bark = width of 1 critical band

$$\text{Bark} = \begin{cases} f/100, & f \le 500\text{Hz} \\ 9 + 4\log_2(f/1000), & f > 500\text{Hz} \end{cases}$$

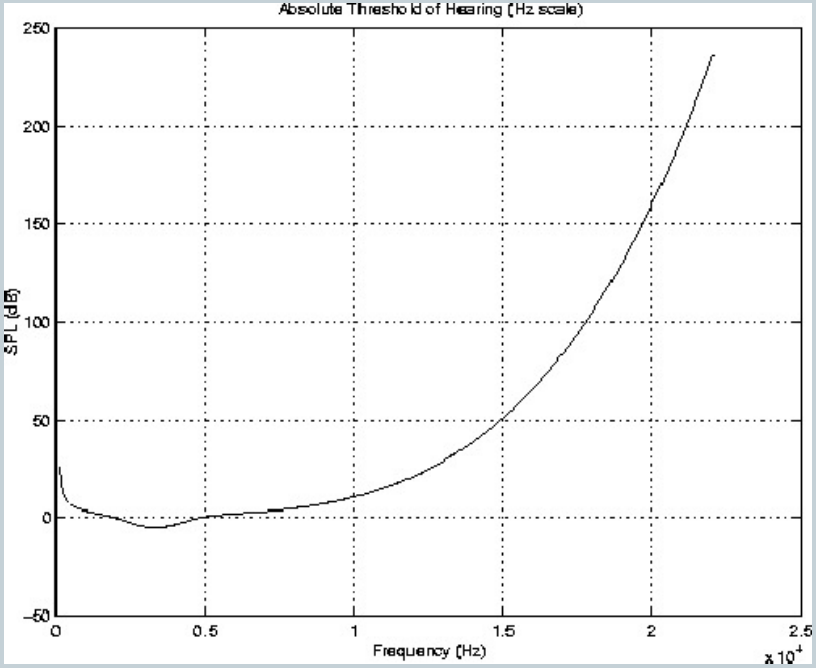


Psychoacoustic Model: Absolute Threshold of Hearing

$$ATHq(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4 (\text{dBSPL})$$

Experiment:

Put a person in a quiet room. Raise level of 1 kHz tone until just barely audible. Vary the frequency, plot:



Observations:

- Ο άνθρωπος είναι ευαίσθητος σε ορισμένες συχνότητες (π.χ. συχνότητες ομιλίας 2-4KHz).
- Οι συχνότητες εντός του "κρίσιμου εύρους ζώνης" (δηλ. το εύρος συχνοτήτων κατά τις οποίες το SNR κάλυψης παραμένει περισσότερο ή λιγότερο σταθερό) αντιμετωπίζονται με τον ίδιο τρόπο.

Psychoacoustic Model
The Weakness of the Human Ear

- Frequency dependent resolution:
  - **Δεν έχουμε τη δυνατότητα να διακρίνουμε μικρές διαφορές στη συχνότητα εντός των κρίσιμων ζωνών** (critical bands).

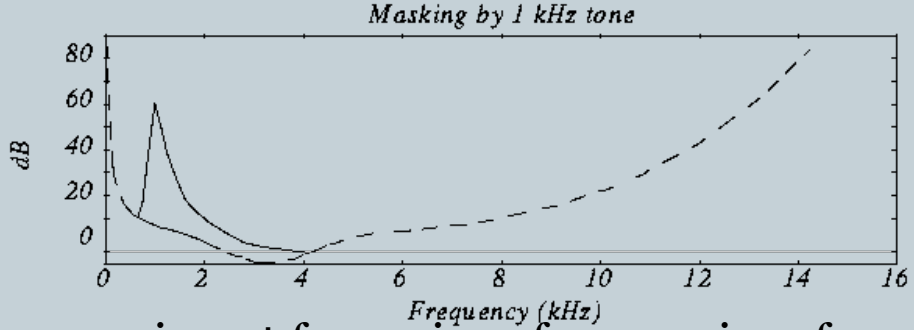
- Auditory masking:
  - Όταν υπάρχουν δύο σήματα τα οποία είναι πολύ κοντά στη συχνότητα, τότε αυτό με πιο δυνατή ένταση θα καλύψει το άλλο.
  - Το ίδιο σήμα για να μπορέσουμε να το ακούσουμε, δηλ. το σήμα το οποίο και κανονικά καλύπτεται, θα πρέπει να είναι πιο δυνατό από **κάποιο συγκεκριμένο κατώφλι** → μας δίνει «χώρο» στην κωδικοποίηση να εισάγουμε «ακουστικό θόρυβο», τον οποίο όμως δεν ακούμε.

Βασικός κανόνας του perceptual coding: χρησιμοποιούμε το μοντέλο σκίασης για να καθορίσουμε τον ελάχιστο αριθμό των bits που χρειάζονται σε κάθε συχνοτική μπάντα έτσι ώστε ο θόρυβος του κβαντιστή να μην είναι αντιληπτός.

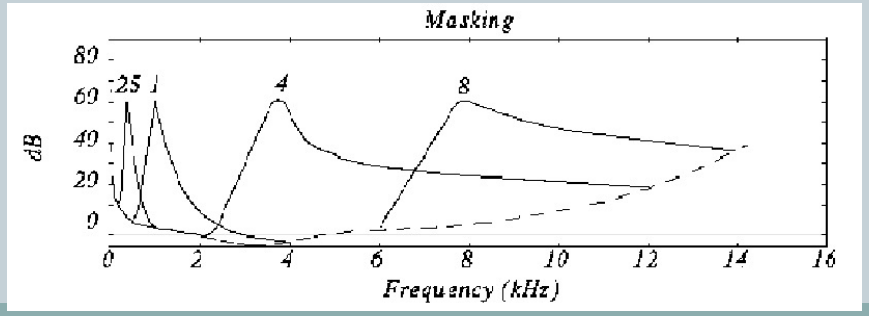
Frequency Masking

Οι άνθρωποι δεν έχουν τη δυνατότητα να ακούνε μικρές διαφορές στη συχνότητα. Π.χ. είναι πολύ δύσκολο να διακρίνουμε ένα σήμα 1,000 Hz από ένα που είναι 1,001 Hz. Αυτό γίνεται ακόμη πιο δύσκολο αν τα δύο σήματα παίζουν ταυτόχρονα

Experiment: Play 1 kHz tone (masking tone) at fixed level (60 dB).
Play test tone at a different level (e.g., 1.1 kHz), and raise level until just distinguishable.

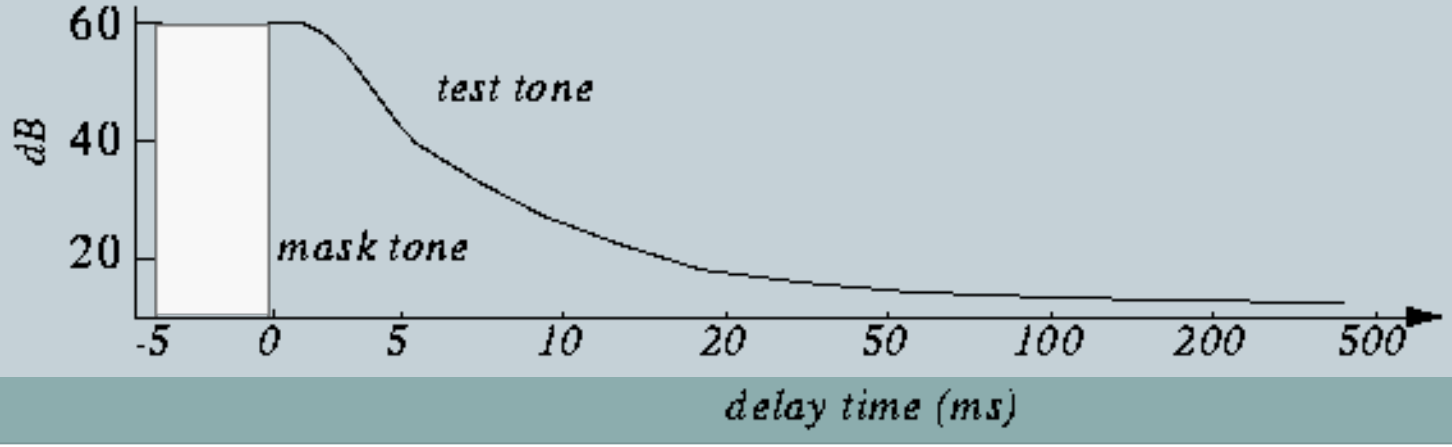


Repeat the previous experiment for various frequencies of masking tones yields:



Temporal Masking

- Ακούμε έναν δυνατό ήχο: από την στιγμή που ο ήχος θα σταματήσει, θα χρειαστούμε κάποιο χρόνο για να ακούσουμε έναν ήχο με χαμηλότερη ένταση.
- Experiment: Play 1 kHz *masking tone* at 60 dB, plus a *test tone* at 1.1 kHz at 40 dB. Test tone can't be heard (it's masked). Stop masking tone, then stop test tone after a short delay.
- Adjust delay time to the shortest time when test tone can be heard (e.g., 5 ms).
- Repeat with different level of the test tone and plot:



MPEG-I Psychoacoustic Models

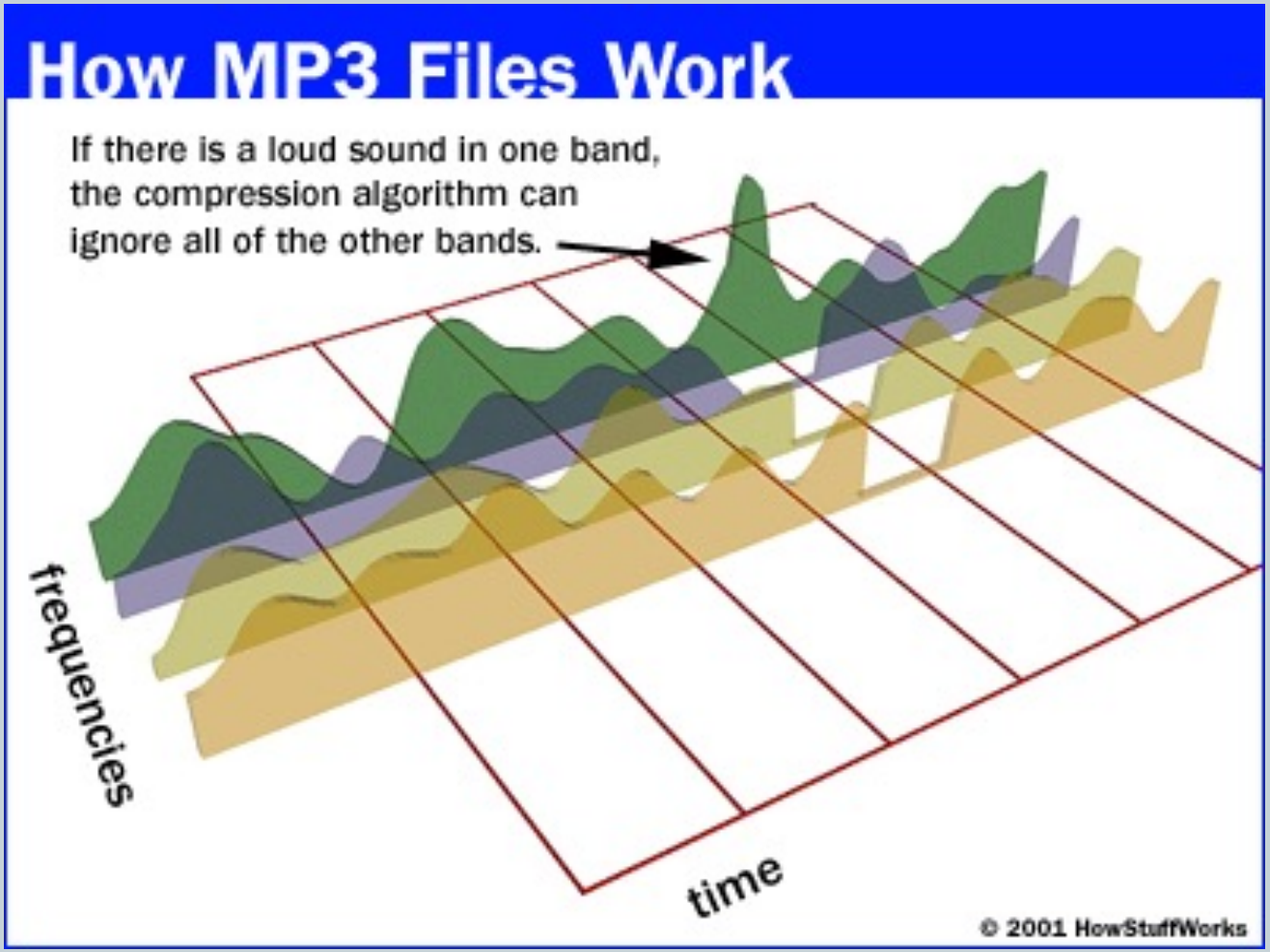
- Calculate a masking threshold for each subband in the filterbank

MPEG-I standard defines two models:

- Psychoacoustic Model 1:
  - Less computationally expensive
  - Makes some serious compromises in what it assumes a listener cannot hear
  - Selects minima of masking threshold values in range of each subband
  - Inaccurate at higher frequencies – recall how subbands are linearly distributed, critical bands are NOT!

- Psychoacoustic Model 2:
  - Provides more features suited for Layer III coding, assuming of course, increased processor bandwidth.
  - If subband wider than critical band:
    - Use minimal masking threshold in subband
  - If critical band wider than subband:
    - Use average masking threshold in subband

MP3 example


How MP3 Files Work
If there is a loud sound in one band, the compression algorithm can ignore all of the other bands.
frequencies
time
© 2001 HowStuffWorks

Masking and Quantization (Example)

- Say, performing the sub-band filtering step on the input results in the following values (for demonstration, see only the first 16/32 bands):

Band	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Level	0	8	12	10	6	2	10	60	35	20	15	2	3	5	3	1

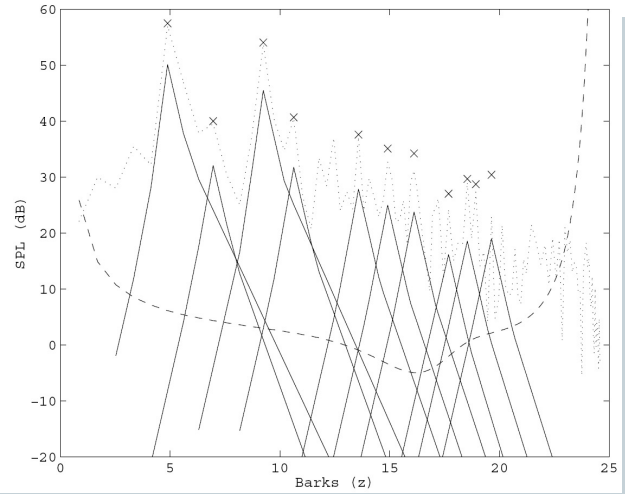
- The 60dB level of the 8th band gives a masking of 12 dB in the 7th band, 15 db in the 9th. (according to the Psychoacoustic model)
- The level 7th band is 10 dB (<12 dB), so ignore it.
- The level in 9th band is 35 dB (>15 dB), so send it.
  - We only send the amount above the masking level
  - Therefore, instead of using 6 bits to encode it, we can use 4 bits – a saving of 2 bits (=12 dB)

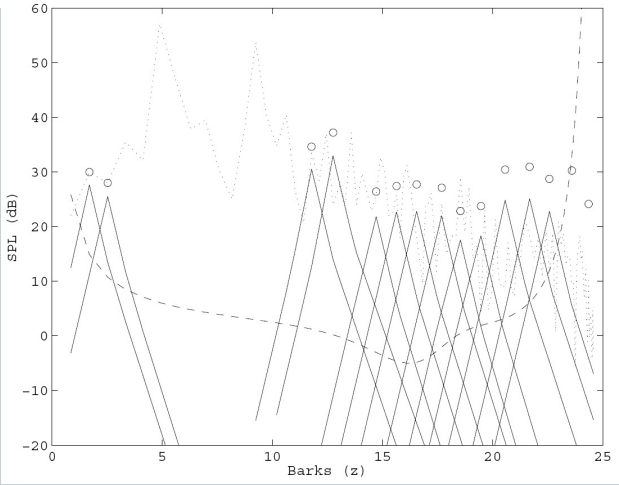
Psychoacoustic Model (Lab)

- Convert samples to frequency domain
  - Use a Hann window and then
  - Use 512-point FFT (Layer I) or 1024 (Layers II and III) sample window.
  - Identification of Tonal and Noise Maskers.
    - Calculation of Individual Masking Thresholds.
  - Calculation of Global Masking Thresholds, (sum of individual masking thresholds and the absolute masking/hearing threshold).

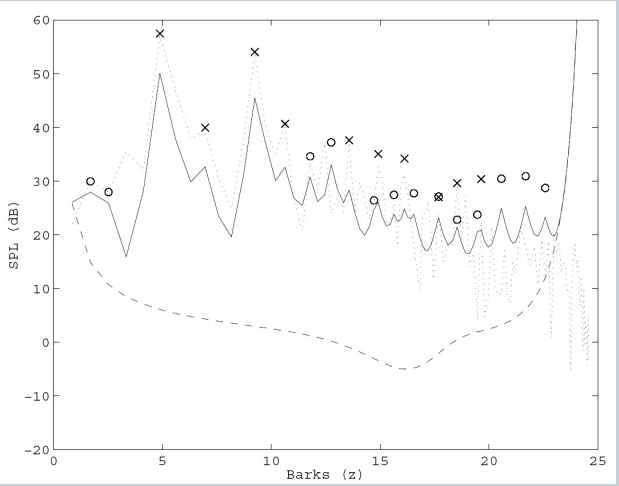
SMR ➔ ratio of the max signal energy level and the min value of the global masking threshold at the given subband.

Demo: Masking Thresholds





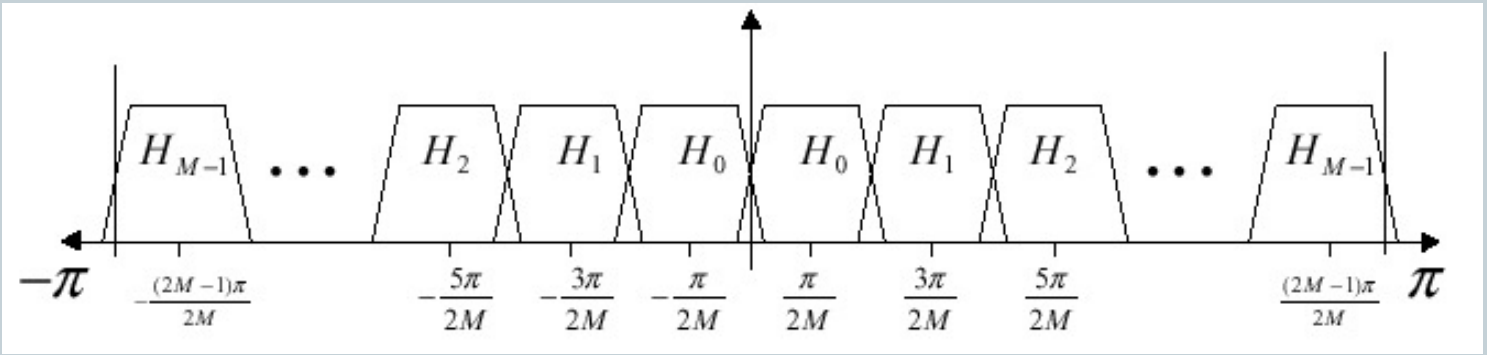
Spreading functions: associated with each of the individual (left) tonal maskers, (right) noise maskers.



Global masking threshold by combining the individual thresholds.

Figures from : T. Painter and A. Spanias, “Perceptual Coding of Digital Audio”, Proceedings of the IEEE 88, 451-513, 2000.

Filterbanks



- Array of bandpass filters covering the entire spectrum
- Break up signal into frequency subbands
  - useful since some frequencies more important than others
- Magnitudes at these important frequencies need to be coded with a fine resolution, thus many bits to encode those, less bits for the frequencies that are not important
- Allows for dynamic allocation of bits (variable coding scheme) to subband or transform coefficients (identifies the perceptual irrelevancies, control the quantization noise)

Analysis and Synthesis Banks

1) Analysis filters divide up the signal
2) Down-sample
3) Quantize
4) Up-sample
5) Synthesis filters remove distortions
6) Reconstruct the signal

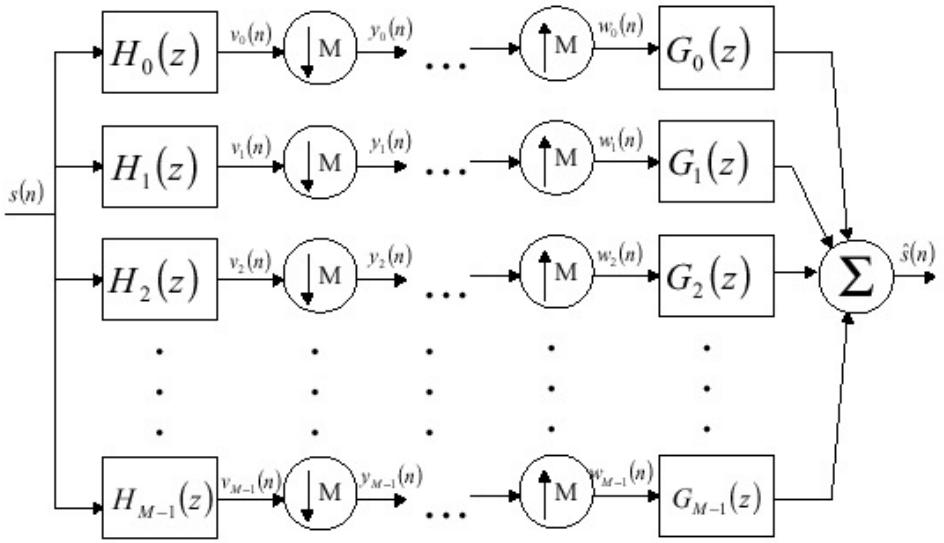


Fig. Uniform M-Band Maximally Decimated Analysis-Synthesis Filterbank

Filterbank Implementation

- Cosine Modulated Perfect reconstruction (PR) M-Bandbanks with 32 filters each & Modified Discrete Cosine Transform (MDCT)

$$h_k(n) = w(n)\sqrt{\frac{2}{M}}\cos[\frac{(2n+M+1)(2k+1)\pi}{4M}]$$

where w(n) is a lowpass filter $w(n) = \sin[(n+\frac{1}{2})\frac{\pi}{2M}],\ M = 32$

Synthesis filter: $g_k(n) = h_k(2M-1-n)$

- The output is a delayed version of the input (linear phase)
- Distortion arises from quantization only

Quantization

- Quantization Noise is the difference between the analog signal and the digital representation, and arises as a result of the error in the quantization of the analog signal.

  N Bits ➔ 2^N levels

- With each increase in the bit level, the digital representation of the analog signal increases in fidelity, and the quantization noise becomes smaller.

- In our case: The number of quantizer levels for each spectral component is obtained from a dynamic bit allocation rule that is controlled by a psychoacoustic model

  bits = $\log_2(2^N/\min(\text{masking threshold}))$ for each subband

References

- T. Painter and A. Spanias, “Perceptual Coding of Digital Audio”, Proceedings of the IEEE 88, 451-513, 2000.
- E.D. Scheirer, “The MPEG-4 Structured Audio”, ICASSP-98.
- D. Pan, “A Tutorial on MPEG/Audio Compression”, *IEEE Multimedia Journal,* 1995. http://www.ee.columbia.edu/~dpwe/e6820/papers/Pan95-mpega.pdf.
- R. Raissi, The Theory Behind MP3, http://www.mp3-tech.org/programmer/docs/mp3_theory.pdf
- P. Noll, MPEG Digital Audio Coding Standards, Chapter in: IEEE Press/CRC Press "The Digital Signal Processing Handbook”(ed.: V.K. Madisettiand D. B. Williams), pp. 40-1 - 40-28, 1998.
- Perceptual Coding: How Mp3 Compression Works, Exploration: https://www.soundonsound.com/techniques/perceptual-coding-how-mp3-compression-works