

Perceptual Coding of Audio Signals – A Tutorial

Copyright 1993, 1995,
1998, 2001,2003

James D. Johnston
now at Microsoft Corporation

What is Coding for?

Coding, in the sense used here, is the process of reducing the bit rate of a digital signal.

The coder input is a digital signal.

The coder output is a smaller (lower rate) digital signal.

The decoder reverses the process and provides (an approximation to) the original digital signal.

Historical Coder “Divisions”:

Lossless Coders

vs.

Lossy Coders

Or

Numerical Coders

vs.

Source Coders

Lossless Coding:

Lossless Coding commonly refers to coding methods that are completely reversible, i.e. coders wherein the original signal can be reconstructed bit for bit.

Lossy Coding:

Lossy coding commonly refers to coders that create an approximate reproduction of their input signal. The nature of the loss depends entirely on the kind of lossy coding used.

Source Coding:

Source Coding can be either lossless or lossy.

In most cases, source coders are deliberately lossy coders, *however*, this is not a restriction on the method of source coding. Source coders of a non-lossy nature have been proposed for some purposes.

Source Coding:

Removes redundancies through estimating a model of the source generation mechanism. This model may be explicit, as in an LPC speech model, or mathematical in nature, such as the "transform gain" that occurs when a transform or filterbank diagonalizes the signal.

Source Coding:

Typically, the source coder uses the source model to increase the SNR or reduce another error metric of the signal by the appropriate use of signal models and mathematical redundancies.

Typical Source Coding Methods:

LPC analysis (including
dpcm and its derivatives
and enhancements)

Sub-band Coding

Transform Coding

Multipulse Analysis by
Synthesis

Vector Quantization

This list is not exhaustive

Well Known Source Coding Algorithms:

Delta Modulation

G728

DCPM

LDCELP

ADPCM

LPC-10E

G721

Numerical Coding:

Numerical coding is almost always a lossless type of coding. Numerical coding, in its typical usage, means a coding method that uses abstract numerical methods to remove redundancies from the coded data.

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

Common Numerical Coding Techniques:

Huffman Coding

Arithmetic Coding

Ziv-Lempel (LZW) Coding

This list is not exhaustive

Numerical Coding (cont.):

Typically, numerical coders use “entropy coding” based methods to reduce the actual bit rate of the signal.

Source coders most often use signal models to reduce the signal redundancy, and produce lossy coding systems.

Both methods work by considering the source behavior.

Both methods attempt to reduce the Redundancy of the original signal.

Perceptual Coding:

Perceptual coding uses a model of the destination, i.e. the human being who will be using the data, rather than a model of the signal source.

Perceptual coding attempts to remove parts of the signal that the human cannot perceive.

Perceptual Coding (cont.):

Is a *lossy* coding method.

The imperceptible information removed by the perceptual coder is called the

irrelevancy

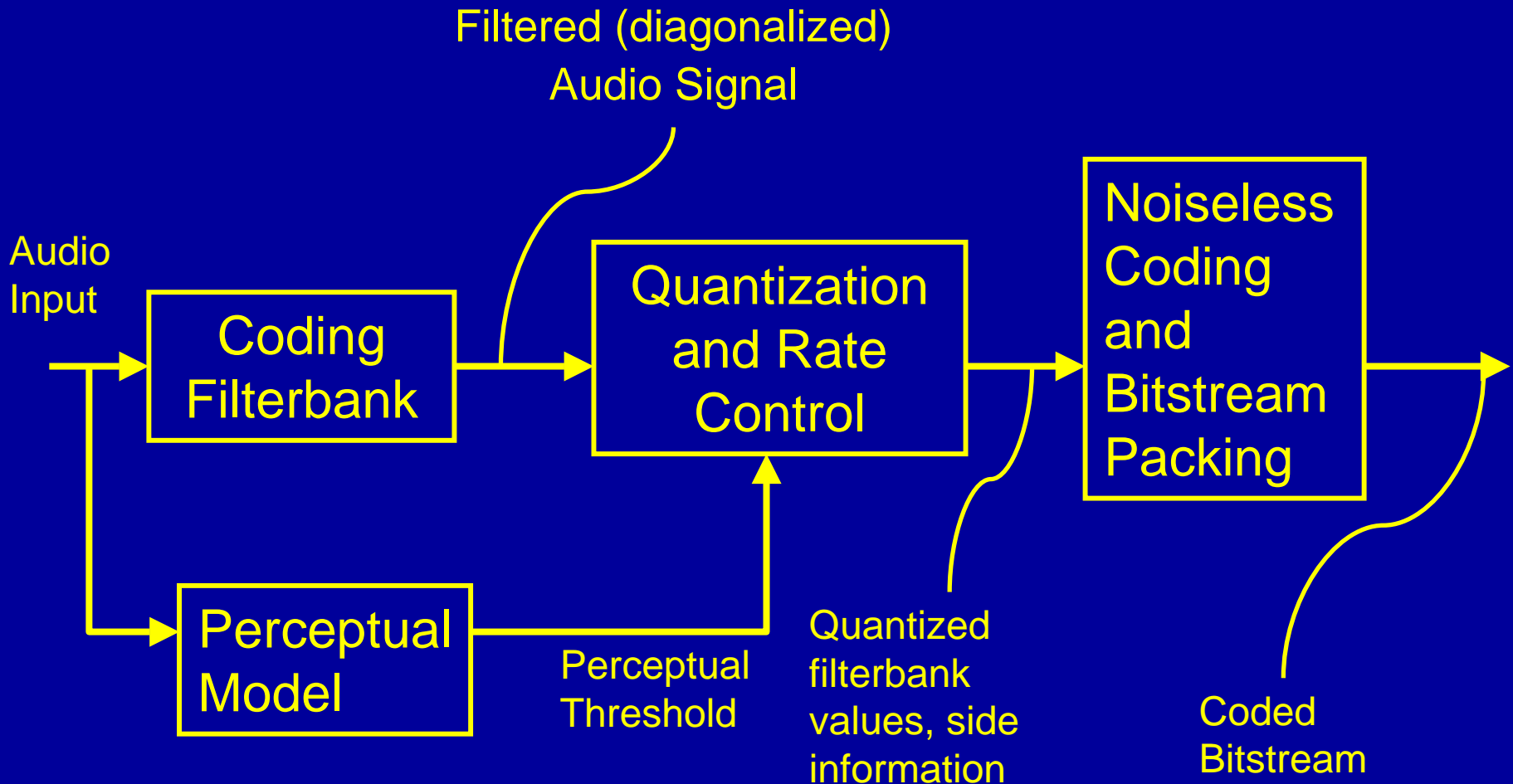
of the signal.

In practice, most perceptual coders attempt to remove both ***irrelevancy*** and ***redundancy*** in order to make a coder that provides the lowest bit rate possible for a give audible quality.

Perceptual Coding (cont.):

Perceptual coders will, in general, have a *lower* SNR than a source coder, and a higher perceived quality than a source coder of equivalent bit rate.

Perceptual Audio Coder Block Diagram



Auditory Masking Phenomena:

The “Perceptual Model”

What is Auditory Masking:

The Human Auditory System (HAS) has a limited detection ability when a stronger signal occurs near (in frequency and time) to a weaker signal. In many situations, the weaker signal is imperceptible even under ideal listening conditions.

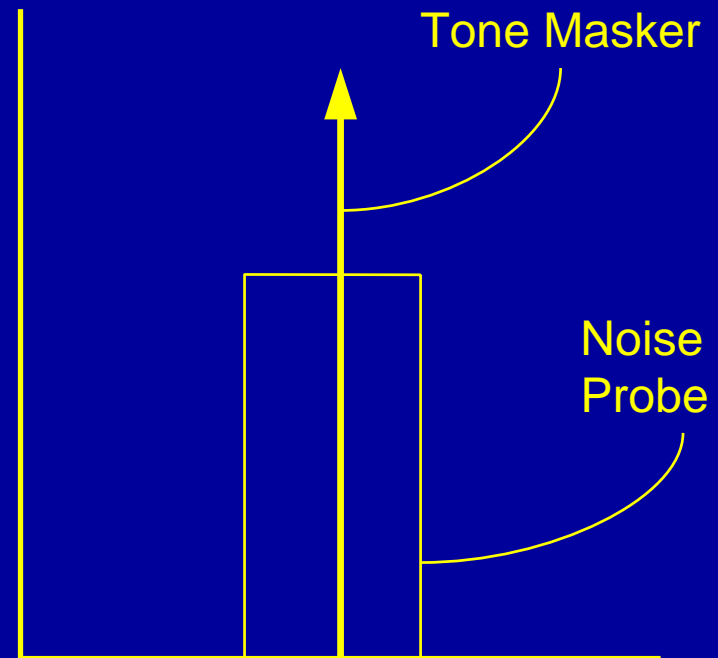
Auditory Masking Phenomena (cont.)

First Observation of Masking:

If we compare:

Tone Masker
to
Tone Masker plus noise

The energy of the 1-bark wide probe is 15.0 dB below the energy of the tone masker.



THE NOISE IS AUDIBLE

Auditory Masking Phenomena (cont.)

The Noise is ***NOT*** Masked!

In this example, a masker to probe ratio of approximately 25 dB will result in complete masking of the probe.

Auditory Masking Phenomena (cont.)

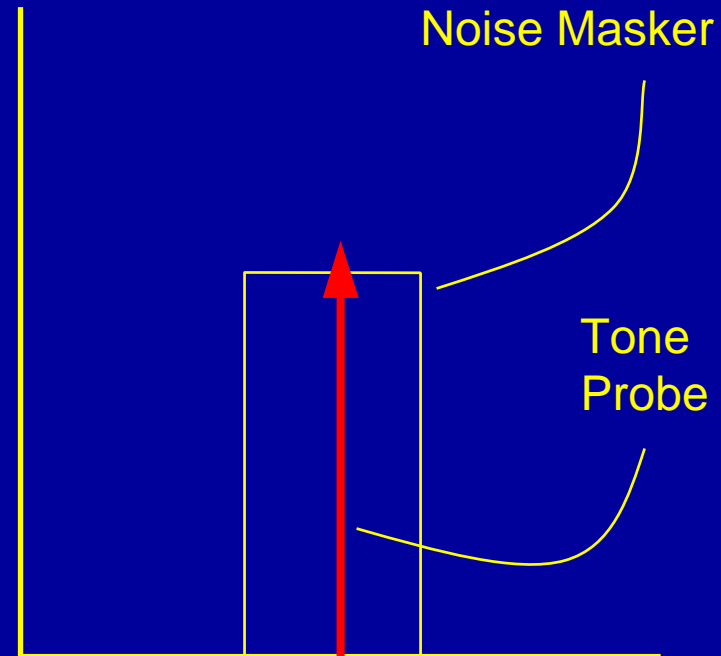
2nd Demonstration of Masking:

If we compare:

Noise Masker
to

Noise Masker plus tone
probe

The energy of the 1-bark wide masker
is 15 dB above the tone probe.



The Tone is NOT Audible

Auditory Masking Phenomena (cont.)

The Tone is ***COMPLETELY*** Masked

In this case, a masker to probe ratio of approximately 5.5 dB will result in complete masking of the tone.

Auditory Masking Phenomena (cont.)

Auditory Masking Phenomena (cont.):

There is an asymmetry in the masking ability of a tone and narrow-band noise, when that noise is within one critical band.

This asymmetry is related to the short-term stability of the signal in a given critical bandwidth.

Critical Bandwidth?

What's this about a *critical bandwidth*?

A critical bandwidth dates back to the experiments of Harvey Fletcher. The term critical bandwidth was coined later. Other people may refer to the “ERB” or equivalent rectangular bandwidth. They are all manifestations of the same thing.

What is that?

Auditory Masking Phenomena (cont.)

A critical band or critical bandwidth

is a range of frequencies over which the masking SNR remains more or less constant.

For example, in the demonstration, any noise signal within $\pm .5$ critical band of the tone will produce nearly the same masking behavior as any other, as long as their energies are the same.

Auditory Masking Phenomena (cont.)

Auditory Filterbank:

The mechanical mechanism in the human cochlea constitute a mechanical filterbank. The shape of the filter at any one position on the cochlea is called the *cochlear filter* for that point on the cochlea. A *critical band* is very close to the passband bandwidth of that filter.

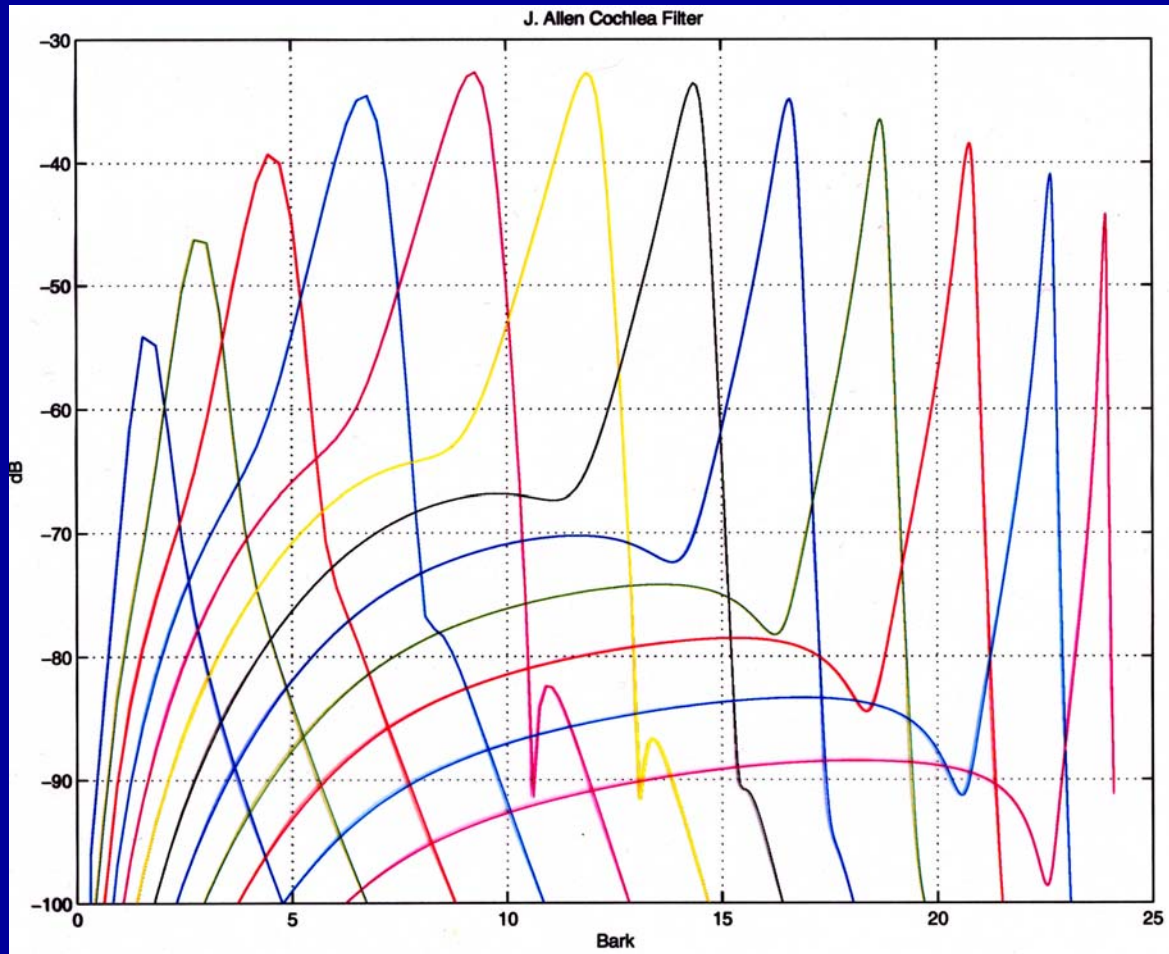
Auditory Masking Phenomena (cont.)

ERB

A newer take on the bandwidth of auditory filters is the “Equivalent Rectangular Bandwidth”. It results in filters slightly narrower at low frequencies, and substantially narrower at mid and high frequencies.

The “ERB scale” is not yet agreed upon.

J. Allen Cochlea Filters

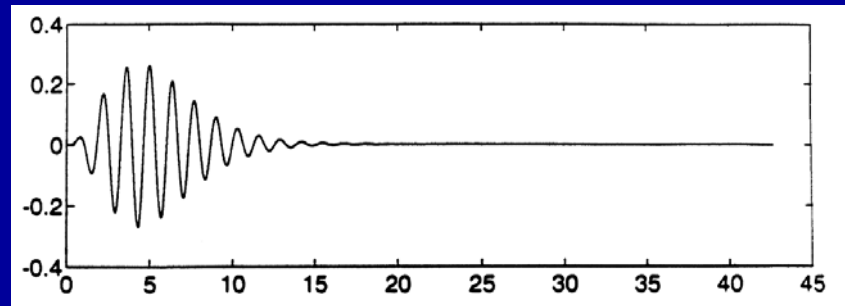


Copyright 1993, 1995,
1998, 2001, 2003

James D. Johnston
now at Microsoft Corporation

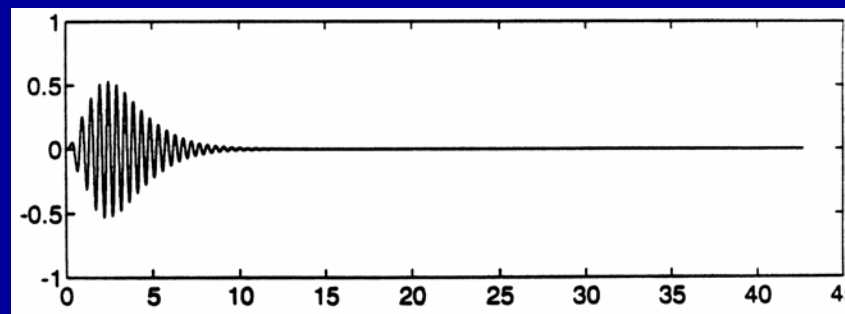
Two Example Cochlear Filters: Time-Domain Response

Impulse
response,
cochlear
filter
centered at
750 Hz



Time (ms)

Impulse
response,
cochlear
filter
centered at
2050 Hz



Time (ms)

Auditory Masking Phenomena (cont.)

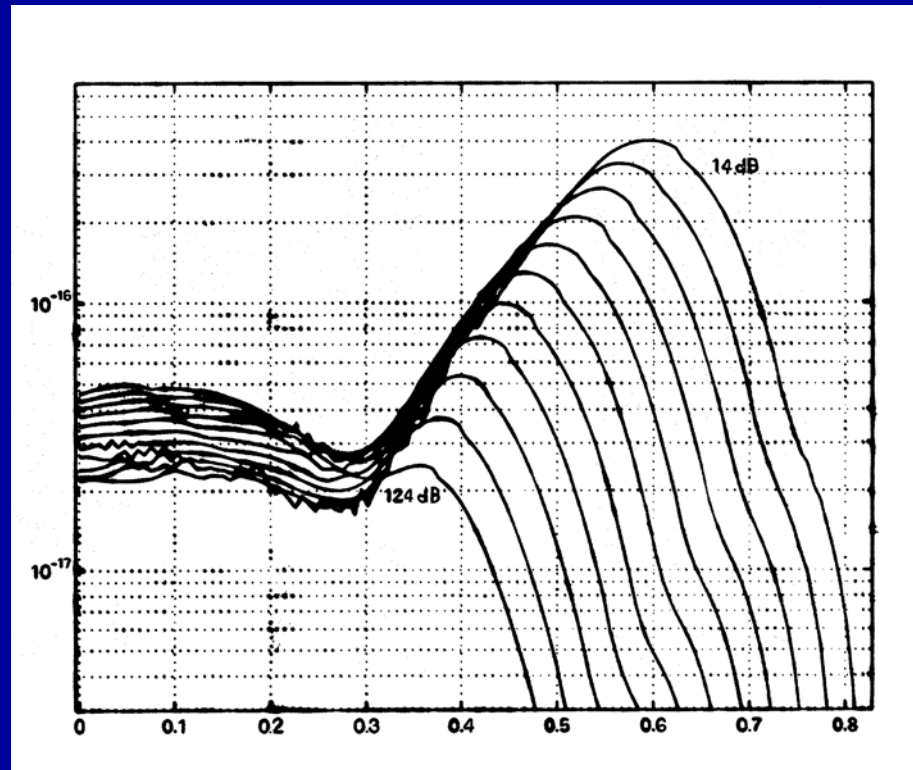
The Cochlear Filterbank

At this time, it seems very likely that the cochlear filterbank consists of two part, a lowpass filter and a highpass filter, and that one filter is tuned via the action of outer hair cells.

This tuning changes the overlap of the two filters and provides both the compression mechanism and the behavior of the upward spread of masking.

Auditory Masking Phenomena (cont.)

Neural Tuning for 5kHz tonal Stimulus (14 – 124 dB SPL)



Distance from Stapes (cm)

Auditory Masking Phenomena (cont.)

The *Bark Scale*

The bark scale is a standardized scale of frequency, where each “Bark” (named after Barkhausen) constitutes one critical bandwidth, as defined in Scharf’s work.

This scale can be described as approximately equal-bandwidth up to 700Hz and approximately 1/3 octave above that point.

Auditory Masking Phenomena (cont.)

Auditory Masking Phenomena (cont.)

A convenient and reasonably accurate approximation for conversion of frequency in Hz to Bark frequency is:

$$B = 13.0 \operatorname{ARCTAN}\left(\frac{0.76f}{1000}\right) + 3.5 \operatorname{ARCTAN}\left(\left(\frac{f}{7500}\right)^2\right)$$

Auditory Masking Phenomena (cont.)

The Bark scale is often used as a frequency scale over which masking phenomenon and the shape of cochlear filters are invariant. While this is not strictly true, this represents a good first approximation.

ERB's Again

The ERB scale appears to provide a more invariant scale for auditory modelling.

With the Bark scale, tone-masking-noise performance varies with frequency.

With a good ERB scale, tone-masking-noise performance is fixed at about 25-30dB.

The Practical Effects of the Cochlear Filterbank in Perceptual Audio Coding:

Describes spreading of masking energy in the frequency domain

Explains the cause of pre-echo and the varying time dependencies in the auditory process

Offers a time/frequency scale over which the time waveform and envelope of the audio signal can be examined in the cochlear domain.

Auditory Masking Phenomena (cont.)

The Spread of Masking in Frequency:

The spread of masking in frequency is currently thought to be due to the contribution of different frequencies to the signal at a given point on the basilar membrane, corresponding to one cochlear filter.

Auditory Masking Phenomena (cont.)

Time vs. Frequency Response of Cochlear Filters

The time extent, or bandwidth, of cochlear filters varies by at least a factor of 10:1 if not more.

As a result, the audio coding filterbank must be able to accommodate changes in resolution of at least 10:1.

Time Considerations in Masking:

Simultaneous Masking

Forward Masking – Masking of a signal by a masker that precedes the masked (probe) signal

Backward Masking – Masking of a probe by a masker that comes after the probe

Forward Masking:

Forward masking of a probe by a masker exists both within the length of the impulse response of the cochlear filter, and beyond that range due to integration time constants in the neural parts of the auditory system.

The length of this masking is $>20\text{ms}$, and is sometimes stated to be as long as several hundred milliseconds. In practice, the decay for post masker masking has two parts, a short *hangover* part and then a longer *decaying* part.

Different coders take advantage of this in different ways.

Limits to Forward Masking

Signals that have a highly coherent envelope across frequency may create low-energy times when coding noise can be unmasked, even when forward masking may be expected to work.

For such signals, Temporal Noise Shaping, (TNS) was developed.

Backward Masking:

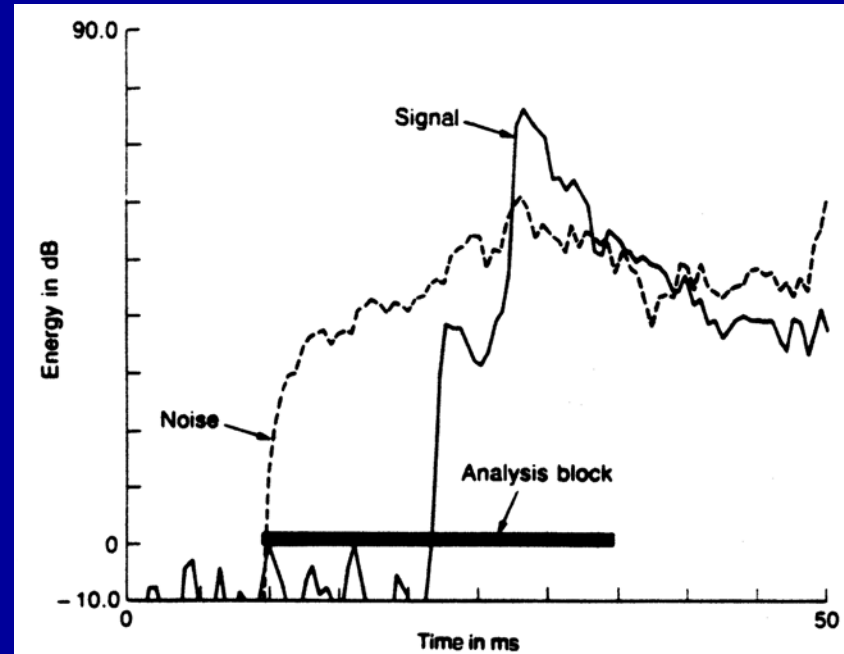
Backward masking appears to be due to the length of the impulse response of the cochlear filter. At high frequencies, backward masking is less than 1ms for a trained subject who is sensitive to monaural time-domain masking effects. Subjects vary significantly in their ability to detect backwardly masked probes.

Effects of the Time Response of the Cochlear Filter on the Coder Filterbank:

The short duration of backward masking is directly opposed to the desire to make the filterbank long in order to extract the signal redundancy. In successful low-rate audio coders, a switched filterbank is a necessity.

The Spread of Masking in Time:

An example of how a filterbank can create a pre-echo.

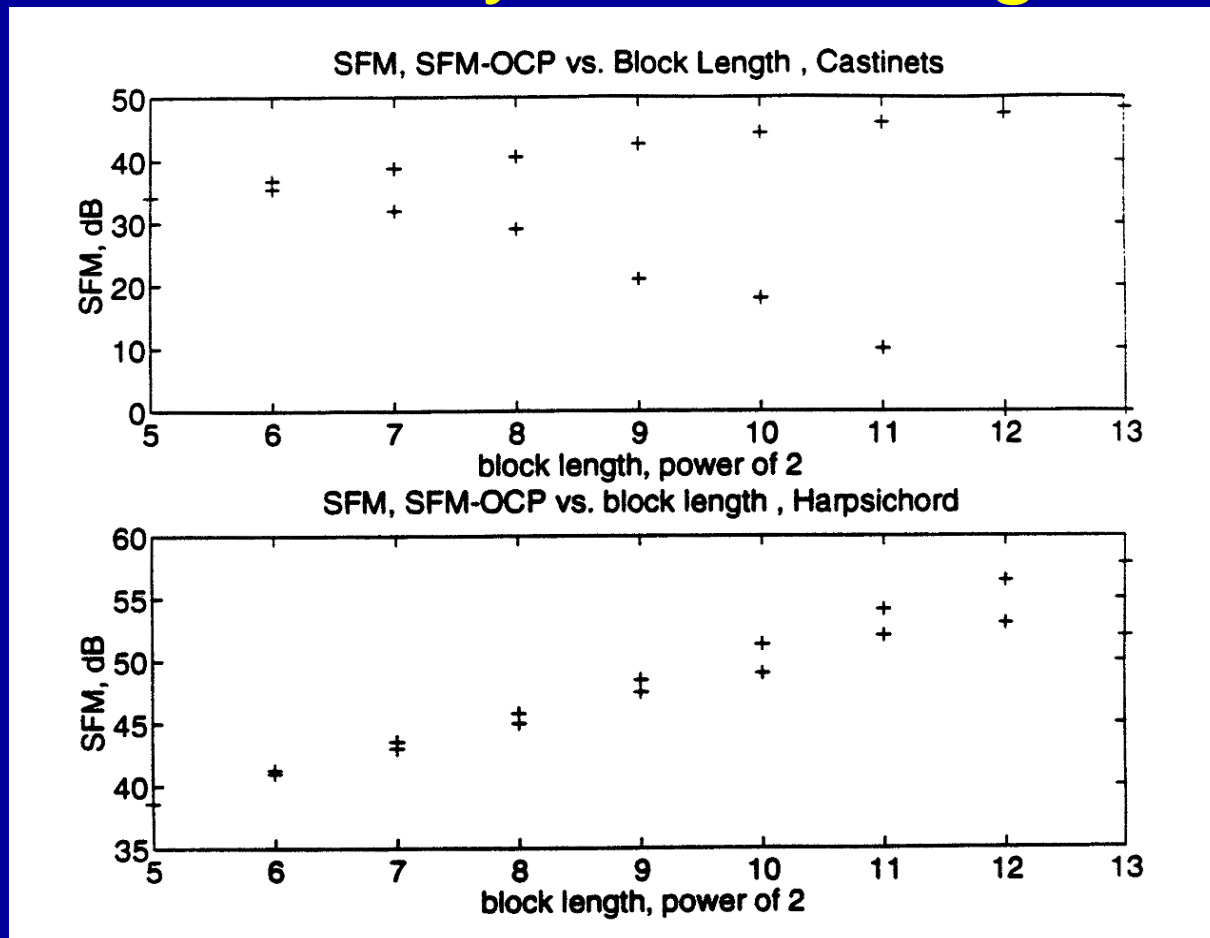


Auditory Masking Phenomena (cont.)

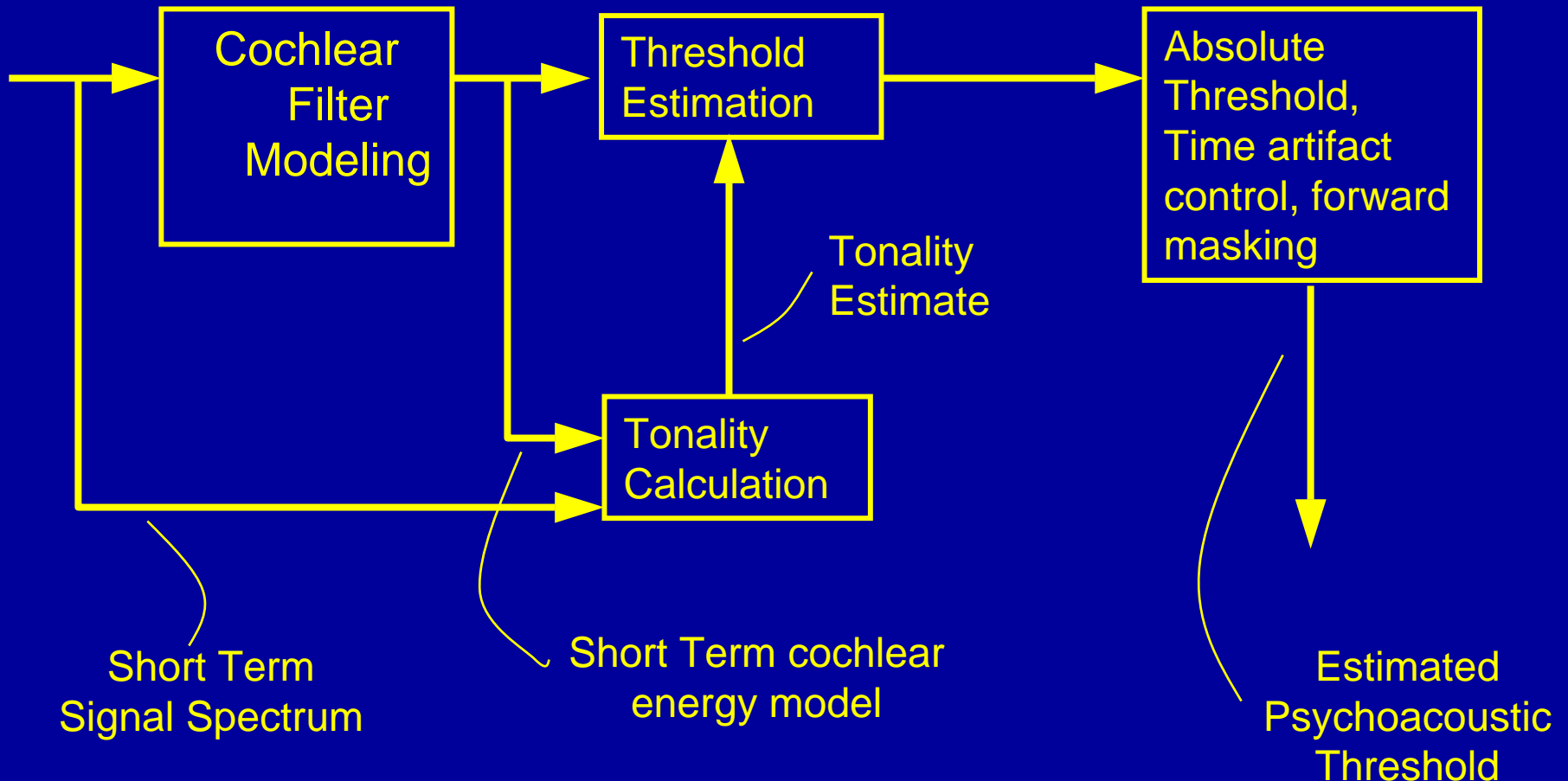
Copyright 1993, 1995,
1998, 2001, 2003

James D. Johnston
now at Microsoft Corporation

An Example of the Tradeoff of Time-Domain Masking Issues vs. Signal Redundancy for Two Signals:



A Typical Psychoacoustic Model:



Issues in Filterbank Design vs. Psychoacoustic Requirements

There are two sets of requirements for filterbank design in perceptual audio coders:

They conflict.

Remember: $ft \geq 1$: The better the frequency resolution, the worse the time resolution.

Requirement 1:

Good Frequency Resolution

Good frequency resolution is necessary to two reasons:

1) Diagonalization of the signal (source coding gain)

And

2) Sufficient frequency resolution to control low-frequency masking artifacts. (The auditory filters are quite narrow at low frequencies, and require good control of noise by the filterbank.)

The Problem with Good Frequency Resolution:

Bad time resolution

Requirement 2:

Good Time Resolution

Good time resolution is necessary for the control of time-related artifacts such as pre-echo and post-echo.

Problems with Good Time Resolution

Not enough signal diagonalization, i.e. not enough redundancy removal.

Not enough frequency control to do efficient coding at low frequencies.

Rule # 2

The filterbank in an audio coder must have both good time resolution ***AND*** good frequency resolution in order to do an efficient job of audio coding.

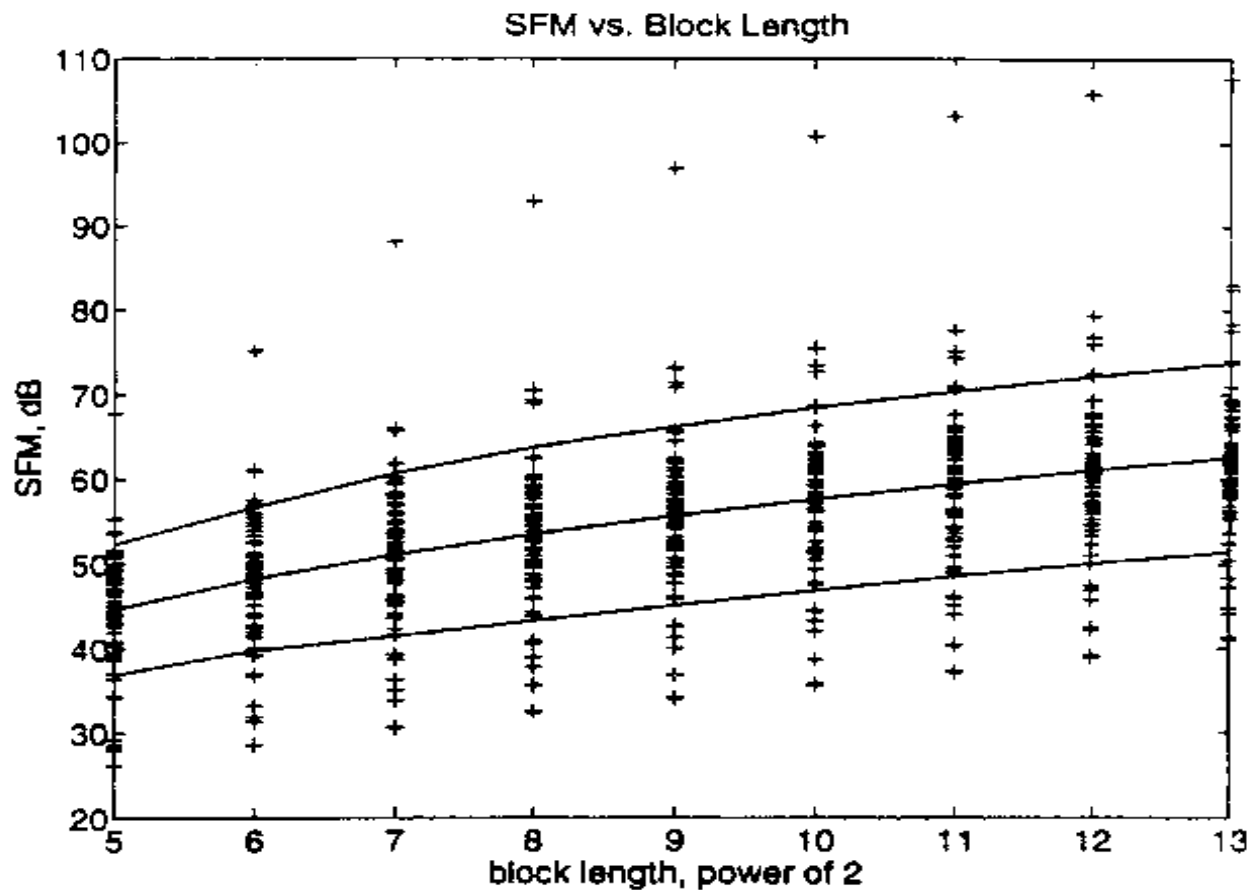
$$\omega t \geq 2\pi$$

Rule #2a

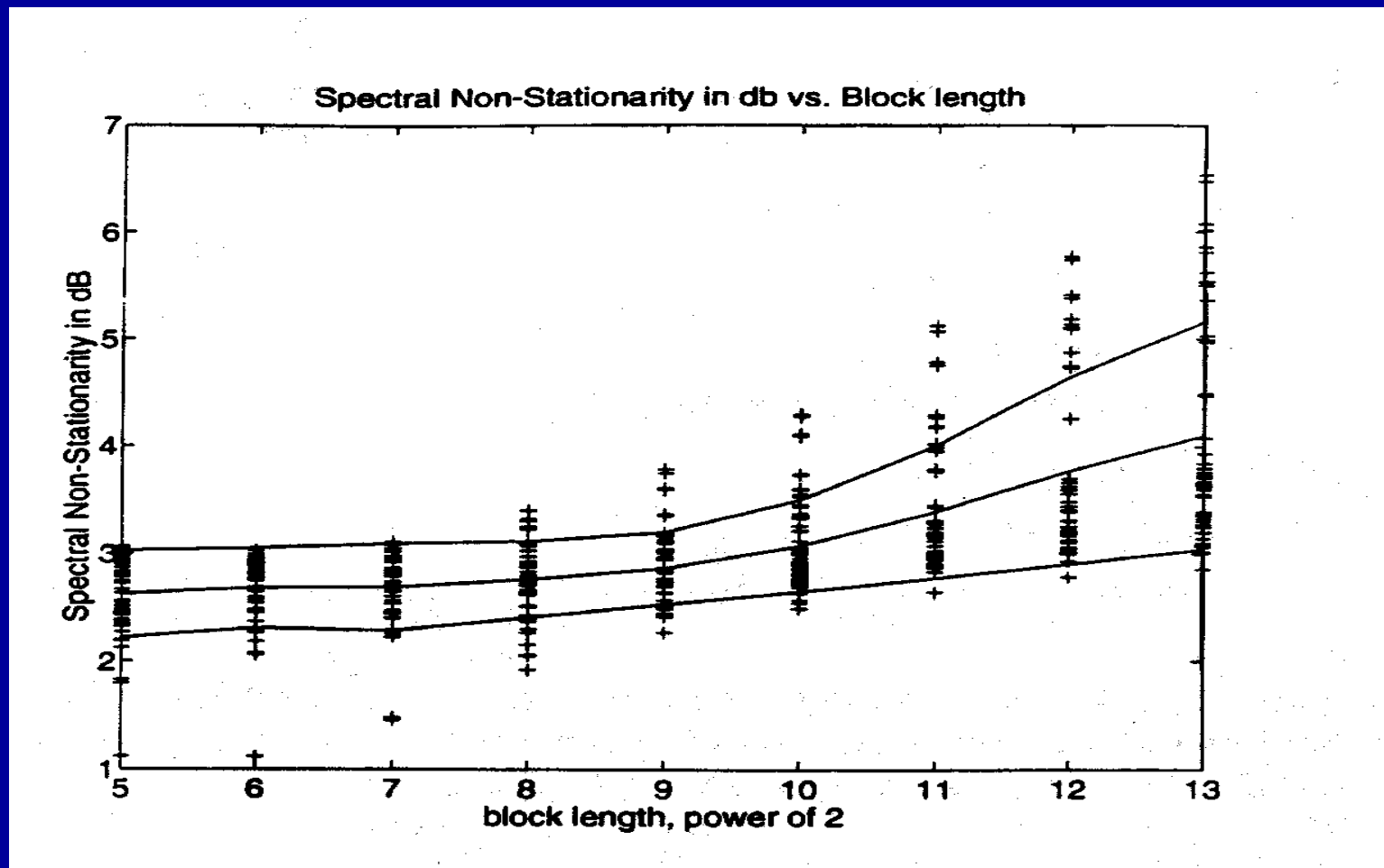
An efficient audio coder must use a time-varying filterbank that responds to both the signal statistics **AND** the perceptual requirements.

Some signal statistics
relevant to audio
coder filterbank
design.

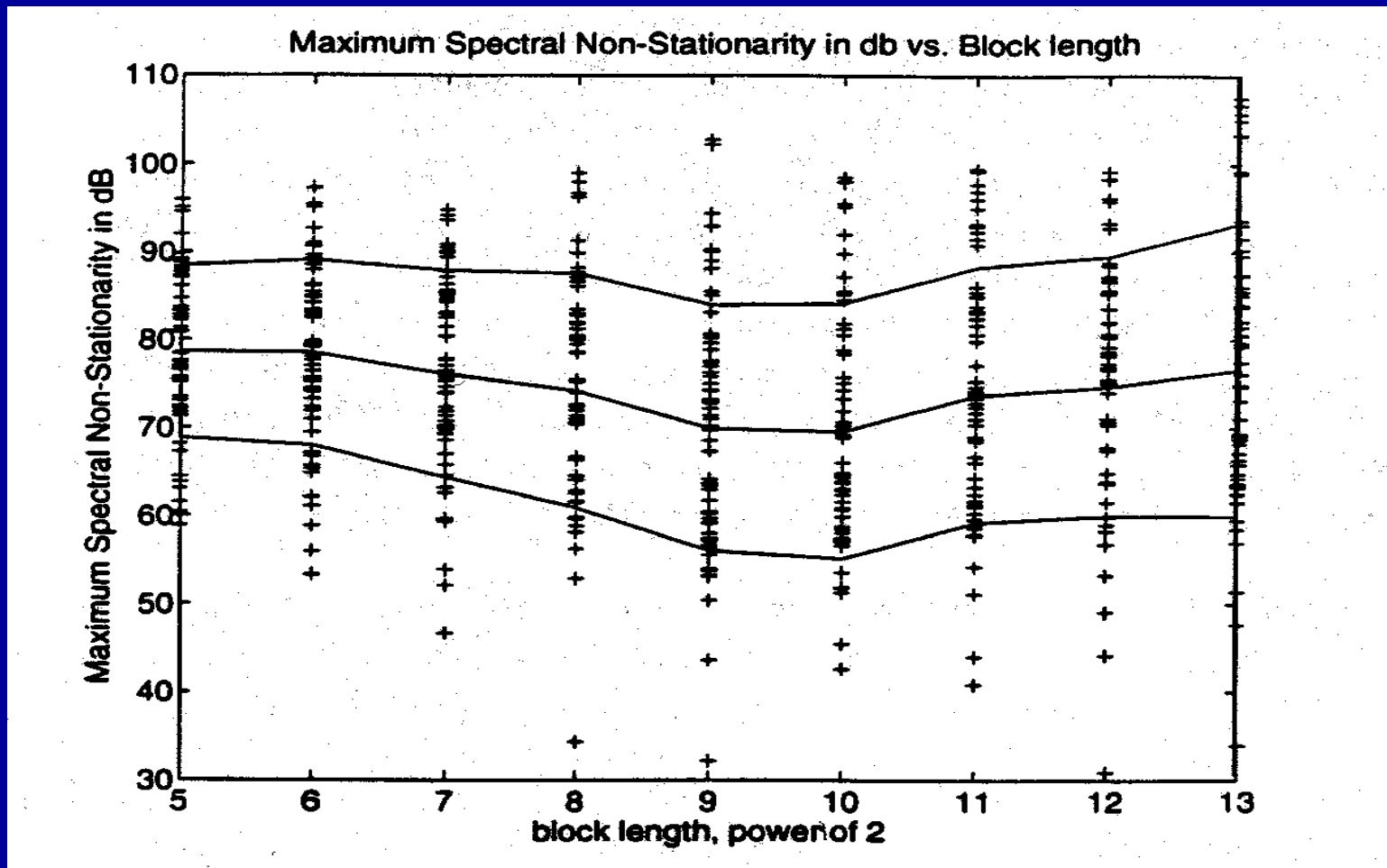
Spectral Flatness Measure as a function of block length



Mean Nonstationarity in Spectrum as a function of block length



Maximum Spectral Nonstationarity



Conclusions about Filterbanks

- 1) A length of about 1024 frequency bins is best for most, if not all, stationary signals.
- 2) A length of 64-128 frequency bins is appropriate for non-stationary signals.

Quantization and Rate Control:

The purpose of the quantization and rate control parts of a perceptual coder is to implement the psychoacoustic threshold to the extent possible while maintaining the required bit rate.

There are many approaches to the quantization and rate control problem. All of them have the same common goals of:

- 1) Enforcing the required rate
- 2) Implementing the psychoacoustic threshold
- 3) Adding noise in less offensive places when there are not enough bits

Quantization and Rate Control

Goals:

Everyone's approach to quantization and rate control is *different*. In practice, one chooses the quantization and rate control parts that interact well with one's perceptual model, bitstream format, and filterbank length(s).

The Use of Noiseless Coding in Perceptual Audio Coders:

There are several characteristics of the quantized values that are obtained from the quantization and rate control part of an efficient perceptual coder.

These Characteristics are:

- 1) The values around zero are the most common values
- 2) The quantizer bins are not equally likely
- 3) In order to prevent the need for sending bit allocation information,
and

in order to prevent the loss of efficiency due to the fact that quantizers do not in general have a number of bins equal to a power of two, some self-termination kind of quantizer value transmission is necessary.

Huffman Codes:

- 1) Are the best-known technique for taking advantage of non-uniform distributions of single tokens.
- 2) Are self-terminating by definition.

Huffman Codes are **NOT** good at:

- 1) Providing efficient compression when there are very few tokens in a codebook.
- 2) Providing efficient compression when there is a relationship between successive tokens.

Arithmetic and LZW coding are good at dealing with symbols that have a highly non-uniform conditional symbol appearance, and with symbols that have a wide probability distribution

but

- 1) They require either extra computation, integer specific programming, or extra RAM in the decoder
- 2) They require a longer training sequence, or a stored codebook corresponding to such a sequence
or
- 3) They have a worse bound on compression efficiency
and
- 4) They create difficulties with error recovery and/or signal break in because of their history dependence,
therefore

the more sophisticated noiseless coding algorithms are not well fitted to the audio coding problem.

The Efficient Solution:

Multi-symbol Huffman codes, i.e. the use of Huffman codes where more than one symbol is included in one Huffman codeword.

Such codebooks eliminate the problems inherent with “too small” codebooks, take a limited advantage of inter-symbol correlation, and do not introduce the problems of history or training time.

An Example Codebook Structure:

The MPEG-AAC Codebook Structure

Codebook Number	Largest Absolute Value	Codebook Dimension	Signed or Unsigned
0	0	*	*
1	1	4	s
2	1	4	s
3	2	4	u
4	2	4	u
5	3	2	s
6	3	2	s
7	7	2	u
8	7	2	u
9	12	2	u
10	12	2	u
11	16 (esc)	2	u

The Problem of Stereo Coding:

There are several new issues introduced when the issue of stereophonic reproduction is introduced:

- 1) The problem of Binarual Masking Level Depression
- 2) The problem of image distortion or elimination

What is Binaural Masking Level Depression (BLMD)?:

At lower frequencies, <3000 Hz, the HAS is able to take the phase of interaural signals into account. This can lead to the case where, for instance, a noise image and a tone image can be in different places. This can reduce the masking threshold by up to 20dB in extreme cases.

Stereo Coding (cont.):

BMLD can create a situation whereby a signal that was “the same as the original” in a monophonic setting sounds substantially distorted in a stereophonic setting.

Two good, efficient monophonic coders do ***NOT*** make one good efficient stereo coder.

Stereo Coding (cont.):

In addition to BLMD issues, a signal with a distorted high-frequency envelope may sound “transparent” in the monophonic case, but will ***NOT*** in general provide the same imaging effects in the stereophonic case.

BMLD

Both the low-frequency BLMD and the high-frequency envelope effects behave quite similarly in terms of stereo image impairment or noise unmasking, when we consider signal envelope at high frequencies or waveforms themselves at low frequencies. The effect is not as strong between 500Hz and 2 kHz.

Stereo Coding (cont):

In order to control the imaging problems in stereo signals, several methods must be used:

- 1) A psychoacoustic model that takes account of BMLD and envelope issues must be included.
- 2) BMLD is best calculated and handled in the M/S paradigm
- 3) M/S, while very good for some signals, creates either a false noise image or a substantial overcoding requirement for other signals.

M/S Coding

M/S coding is mid/side, or mono/stereo coding, M and S are defined as:

$$M=L+R$$

$$S=L-R$$

The normalization of $\frac{1}{2}$ is usually done on the encoding side. L in this example is the left channel, R the right.

Stereo Coding (cont.):

A good stereo coder uses both M/S and L/R coding methods, as appropriate.

The MPEG-AAC algorithm uses a method whereby the selection of M/S vs. L/R coding is made for each of 49 frequency band in each coding block. Protected thresholds for M, S, L, and R are calculated, and each M/S vs. L/R decision is made by calculating the bit cost of both methods, and choosing the one providing the lowest bit rate.

Stereo Coding (cont.):

An M/S coder provides a great deal of redundancy elimination when signals with strong central images are present, or when signals with a strong “surround” component are present.

Stereo Coding (cont.):

Finally, an M/S coder provides better signal recovery for signals that have “matrixed” information present, by preserving the M and S channels preferentially to the L and R channels when one of M or S has the predominant energy.

What's This About "Intensity Stereo" or the MPEG-1 Layer 1,2 "Joint Stereo Mode"?

Intensity stereo is a method whereby the relative intensities of the L and R channels are used to provide high-frequency imaging information. Usually, one signal (L+R, typically) is sent, with two gains, one for L and one for R.

“Intensity Stereo” (cont.):

“Intensity Stereo” Methods do not guarantee the preservation of the Envelope of the Signal for High Frequencies.

For “lower quality” coding, intensity stereo is a useful alternative to M/S stereo coding,

and

For situations where intensity stereo DOES preserve the high-frequency signal envelope, it is useful for high quality audio coding. Such situations are not as common as one might prefer.

Think of intensity stereo as the coder equivalent of a “pan-pot”.

Temporal Noise Shaping

Temporal Noise Shaping (TNS) can help with preserving the envelope in the case of intensity stereo coding,

HOWEVER

the control of TNS and intensity stereo is not yet well understood.

Stereo Coding (cont.):

Finally, a stereo coder must consider the joint efficiency issues when “block switching” on account of a signal attack. If the attack is present in only one channel, the pre-echo must be prevented, while at the same time maintaining efficient coding for the non-attach channel.

This is a tough problem.

What About Intensity Stereo in the MPEG-AAC Standard?

In the AAC standard, intensity stereo can be activated by using one of the “spare” codebooks. The ability to use M/S or intensity stereo coding as needed, in each coding block, allows for extremely efficient coding of both acoustic and pan-pot stereo signals.

So, what about rate control and all that good stuff?

Because the rates of the M, S, L, and R components vary radically from instant to instant, the only reasonable way to do the rate control and quantization issues is to do an “overall” rate control, hence my unwillingness to say “this is 48 kb/s/channel” as opposed to “this is a 96 kb/s stereo-coded signal.”

The more information that one can put under the rate control mechanism at one instant, the better the coder can cross-allocate information in a perceptually necessary sense, hence the same is true for multi-channel audio signals, or even sets of independent audio signals.

Multichannel Audio Issues:

The issue of multichannel audio is a natural extension of the stereophonic coding methods, in that symmetric pairs must be coded with the same stereophonic imaging concerns, and in that joint allocation across all channels is entirely desirable.

Multichannel (cont.):

There are some problems and techniques unique to the multichannel environment:

- 1) Inter-channel prediction.
- 2) Pre-echo in the multichannel coder
- 3) Time delay to “rear” channels

Inter-channel Prediction:

It is thought that under some circumstances, the use of inter-channel prediction may reduce the bit rate.

To the present, this has not been realized in a published coder due to the delay issues in rear channels and the memory required to realize such inter-channel predictors.

Pre-echo in the Multi-channel Setting:

Due to the delay in signals between channel pairs, it is necessary to provide independent block switching for each channel pair, at least, in order to eliminate situations where enormous over-coding requirements occur due to the need to suppress pre-echos.

Time Delay to the “Rear” Channels:

In multi channel audio, there is often a long time delay to the rear channels. While the problems this introduces have, in a sense, been addressed in the prediction and pre-echo comments, this delay in fact makes “joint” processing of more than channel pairs difficult on many if not all levels.

On the other hand, as this decorrelates the bit-rate demand for front and rear channels, it raises the gain available when all channels are jointly processed in the quantization and rate-control sense.

Multichannel:

On the issue of “Backward Compatibility”, or the ability to either send or derive a stereo mixdown from the multichannel coded signal:

(turn on echoplex)

The use of “Matrix” or “L’,R’” matrixing inside the coder is a

BAD IDEA!

Multichannel:

When the 2-channel mixdown is required, it is better to send it as a separate signal pair, rather than as either a pre- or post-matrixed signal, and allowing the appropriate extra bit rate for the extra two channels.

The same is true for the Monophonic mixdown channel.

Why?

Multichannel:

There are several reasons:

- 1) It is better, from the artist's and producer's point of view, to have separate, and deliberately mixed, 1, 2, and multichannel mixdowns.
- 2) In the process of assuring the quality of either the L or L' channel, whichever is derived, the peak bit rate will be the same or more than that which is required to simply send that channel outright. Further more, in this case, additional decoder complexity and memory will be required.

Using Perceptual Audio Coding:

Perceptual audio coding is intended for final delivery applications. It is not advisable for principle recording of signals, or in cases where the signal will be processed heavily

AFTER

the coding is applied.

Using Perceptual Audio Coding:

Perceptual audio coding is applicable where the signal will NOT be reprocessed, equalized, or otherwise modified before the final delivery to the consumer.

The “Tandeming” or “Multiple Encoding” Problem:

There is a one-word solution to the problem
of using multiple encodings.

DON'T

Multiple Encoding (cont.):

If you are in a situation where you must do multiple encodings:

1) Avoid it to the extent possible
and

2) Use a high bit rate for all but the final
delivery bitstream.

Finally:

Perceptual coding of audio is a very powerful technique for the **final** delivery of audio signals in situations where the delivery bit rate is limited, and/or when the storage space is small.