# Digital Speech and Audio Coding

## Lecture 6

## Sub-band and Transform coding of Audio and Speech

Petr Motlicek and Mathew Magimai Doss

Idiap Research Institute, Martigny

http://www.idiap.ch/

Ecole Polytechnique Fédérale de Lausanne, Switzerland

# Outline

- Analysis-Synthesis framework for M-band filter banks
- QMF, PQMF
- MDCT as filter bank
- DFT and DCT as filter bank
- Pre-echo distortion
- Filter bank and lapped transforms
- Sub-band coding, Channel and Phase vocoders
- Transform Coding
- ...

# Analysis-Synthesis Framework for M-band filter banks

Choosing a proper filter-bank:

- Efficient coding heavily depends on adequaly matching the properties of the analysis filter bank to the characteristics of the input signal.
- Always tradeoff between time and frequency resolution.
- Failure to choose proper filter bank - perceptible artifacts in the output or low coding gain - high bit-rates.

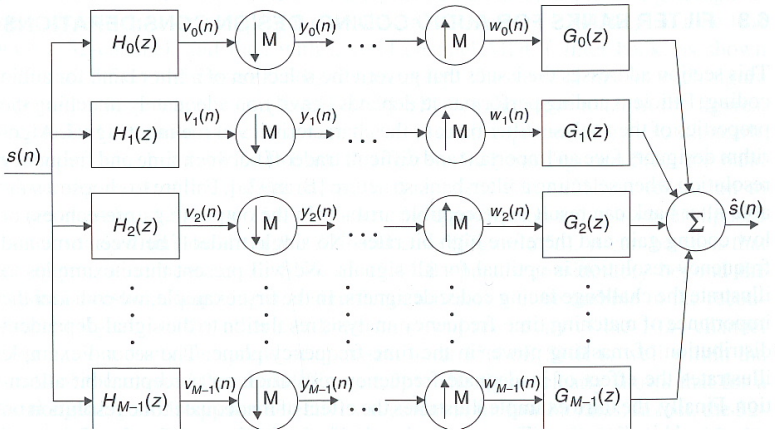


Fig: Uniform M-band maximally decimated analysis-synthesis filter bank (the number of subband samples is equal to the number of input samples.

# Analysis-Synthesis Framework for M-band filter banks

- Unavoidable aliasing between the decimated subband sequences.
- Synthesis filters - removing the imaging distortions introduced by upsampling.
- A-S filters designed to cancel aliasing and imaging distortions according to:

$$\hat{S}(\Omega) = \frac{1}{M} \sum_{k=0}^{M-1} \sum_{l=0}^{M-1} S\Big(\Omega + \frac{2\pi l}{M}\Big) H_k\Big(\Omega + \frac{2\pi l}{M}\Big) G_k(\Omega).$$

- For PR - $\hat{s}(n)$ will be identical to $s(n)$ within a delay, i.e., $\hat{s}(n) = s(n - n_0)$ as long as there is no quantization noise introduced.

# Analysis-Synthesis Framework for M-band filter banks

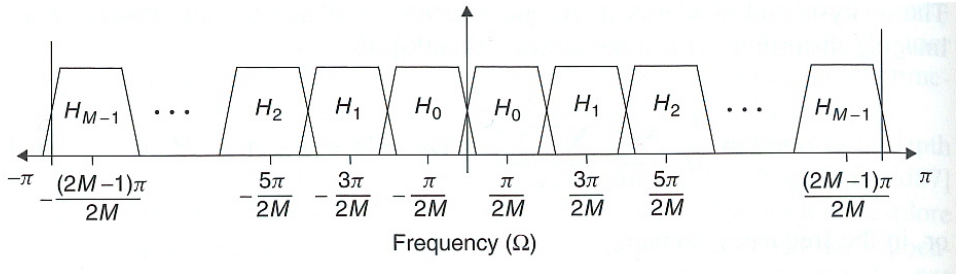


Fig: Magnitude frequency response for a uniform M-band filter bank.

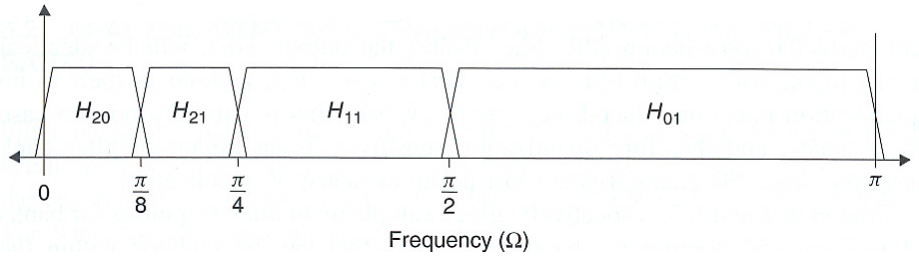


Fig: Magnitude frequency response for an octave-band filter bank.

# Quadrature Mirror Filters (QMFs)

Starting points - two band QMFs - used in early subband algorithms for speech coding and later for first 7kHz standardized wideband audio coding alg (ITU G.722). There is a strong connection between two band PR QMF and discrete wavelet transform.

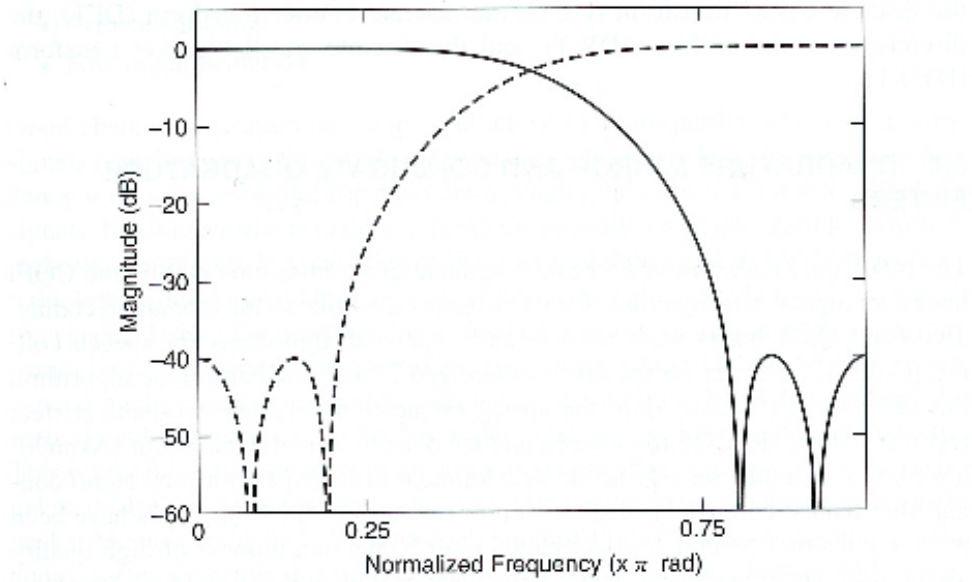


Fig: Two-band Smith-Barnwell CQF filter bank magnitude frequency response with L=8.

# Quadrature Mirror Filters (QMFs)

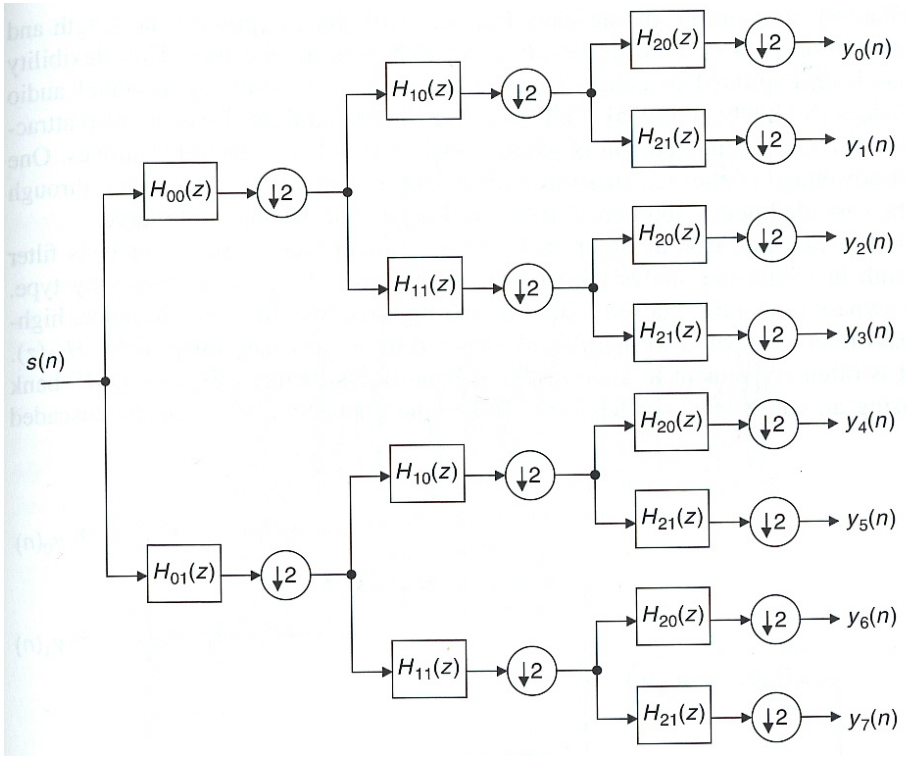


Fig: Tree-structured realization of a uniform 8-channel analysis filter bank.

Advantages - drawbacks:

- Arbitrary partitioning the frequency axis by creating an appropriate cascade - also in nonuniform manner to approximate analysis properties of human ear.

# Quadrature Mirror Filters (QMFs)

- Flexibility to optimize the length and other properties of filters at each node in the tree.
- Computation efficiency.
- Delay - accumulated through the cascaded nodes.

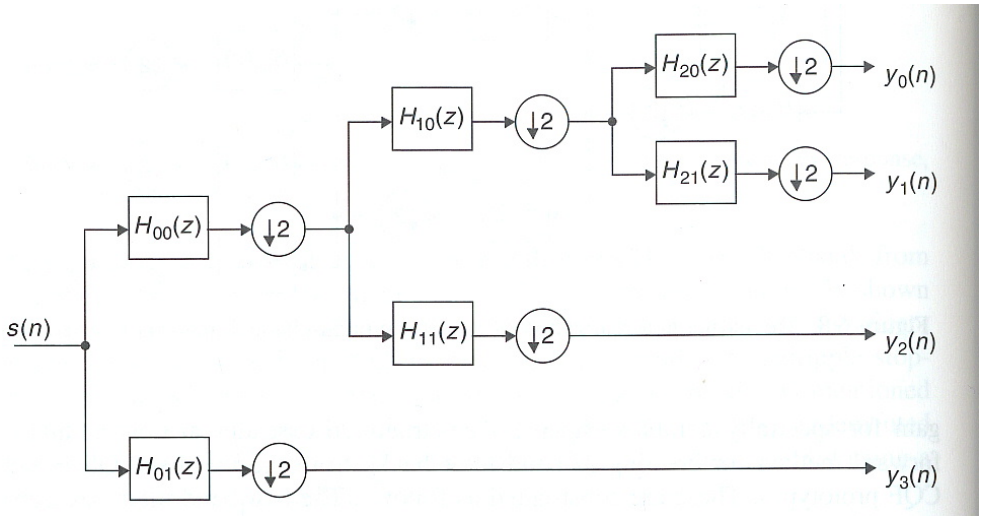


Fig: Tree-structured realization of an octave-band 4-channel analysis filter bank.

# Quadrature Mirror Filters (QMFs)

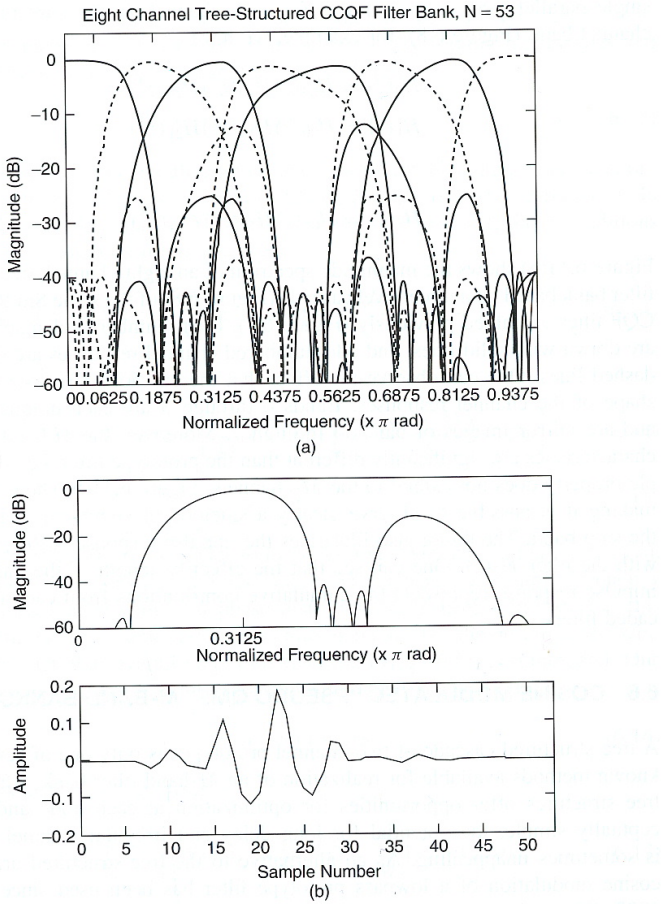


Fig: 8-channel cascaded CQF filter bank - magnitude responses for odd and even-numbered channels. Isolated view of the magnitude freq. response and time-domain impulse response for channel 3. It highlights the presence of the significant side-lobe in the stop-band.

# Cosine Modulated "Pseudo" QMF M-band banks

A tree-structured cascade of 2-channel prototypes is only one of several methods for realization of an M-band filter bank. Cosine modulation of low-pass prototype filter has been used since 1980s to realize parallel M-channel filter bankwith nearly (thats is why pseudo) PR:

- single FIR prototype filter
- uniform linear phase channel response
- overall linear phase - constant group delay
- low complexity
- critical sampling

In PQMF band derivation, phase distortion is completely eliminated from the overall transfer function by forcing analysis and synthesis filter to satisfy mirror image condition: $g_k(n) = h - k(L - 1 - n)$. Adjacent channel aliasing is cancelled by

# Cosine Modulated "Pseudo" QMF M-band banks

establishing precise relation between the analysis and synthesis filters, $H_k(z)$ and $G_k(z)$, respectively. Then filters are given:

$$h_k(n) = 2w(n)cos\Big[\frac{\pi}{M}(k+0.5)\Big(n-\frac{(L-1)}{2}\Big)+\Omega_k\Big]$$

and synthesis filter given by

$$g_k(n) = 2w(n)cos\Big[\frac{\pi}{M}(k+0.5)\Big(n-\frac{(L-1)}{2}\Big)-\Omega_k\Big],$$

where $\Omega_k = (-1)^k\frac{\pi}{4}$. $w(n)$ - L-sample window, a real coefficient liner phase FIR prototype low-pass filter.

# Cosine Modulated "Pseudo" QMF M-band banks

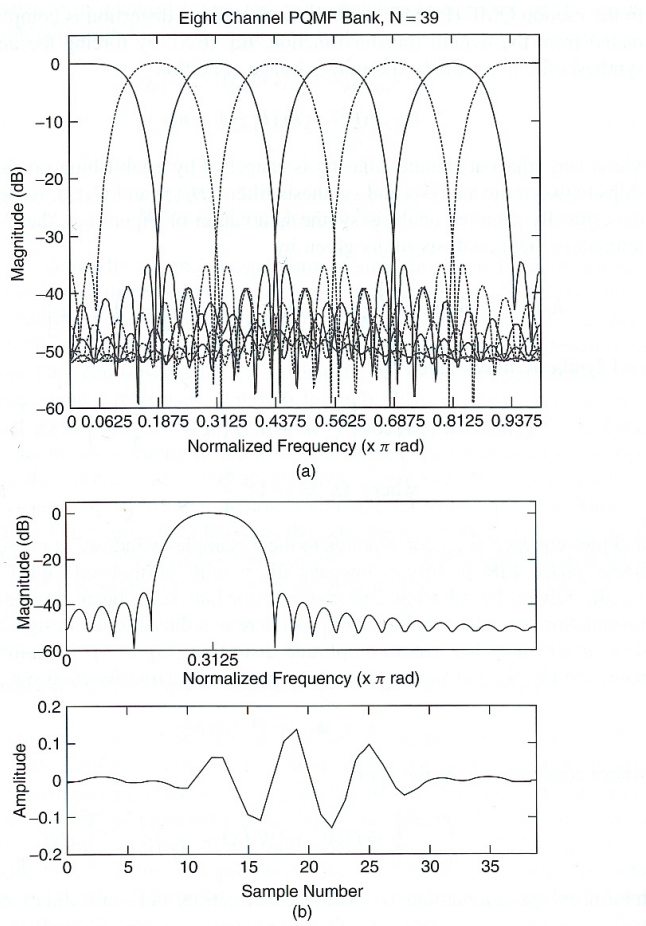


Fig: 8-channel PQMF filter bank - magnitude responses for odd and even-numbered channels. Isolated view of the magnitude freq. response and time-domain impulse response for channel 3. It highlights the presence of the significant side-lobe in the stop-band.

# Cosine Modulated PR M-band banks and MDCT

PQMF used a lot in perceptual audio coding, the overall system must compensate for inherent distortion introduced by lack of PR to avoid audible artifacts. The strategy can be simple - increase prototype filter length - which does not satisfy anyway PR (more preferable due to constraining the sources of output distortion to the quantization stage).

In 1990s, the generalized PR cosine modulated filter banks are possible by appropriately constraining the prototype low-pass filter $w(n)$ and synthesis filters $g_k(n)$.

PR is quaranteed for $h_k(n)$ (given before) if 4 conditions are satisfied:

- $L = 2mM$, where $m$ is integer greater than zero and L is the length of the window $w(n)$.
- $g_k(n) = h_k(L-1-n)$ - time-reversal of the synthesis filter.
- $w(n) = w(L-1-n)$ - FIR low-pass prototype must have linear

# Cosine Modulated PR M-band banks and MDCT

phase.

- Poly-phase components must satisfy the pairwise power complementary requirement.

Let's concentrate mainly on the special case $L = 2M$- development of TDAC filter bank:

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

Forward and Inverse MDCT:

- 50% overlap between blocks - virtually eliminating the blocking artifacts compared to non-overlapped transform coders.
- Despite that - still critically sampled.
- Forward MDCT: $X(k) = \sum_{n=0}^{2M-1} x(n)h_k(n)$ - a series of inner products between the M analysis filter responses $h_k(n)$ and the input.

# Cosine Modulated PR M-band banks and MDCT

- Inverse MDCT: $x(n) = \sum_{n=0}^{M-1}[X(k)h_k(n) + X^P(k)h_k(n+M)]$, where $X^p(k)$ denotes the previous block of transform coefficients.

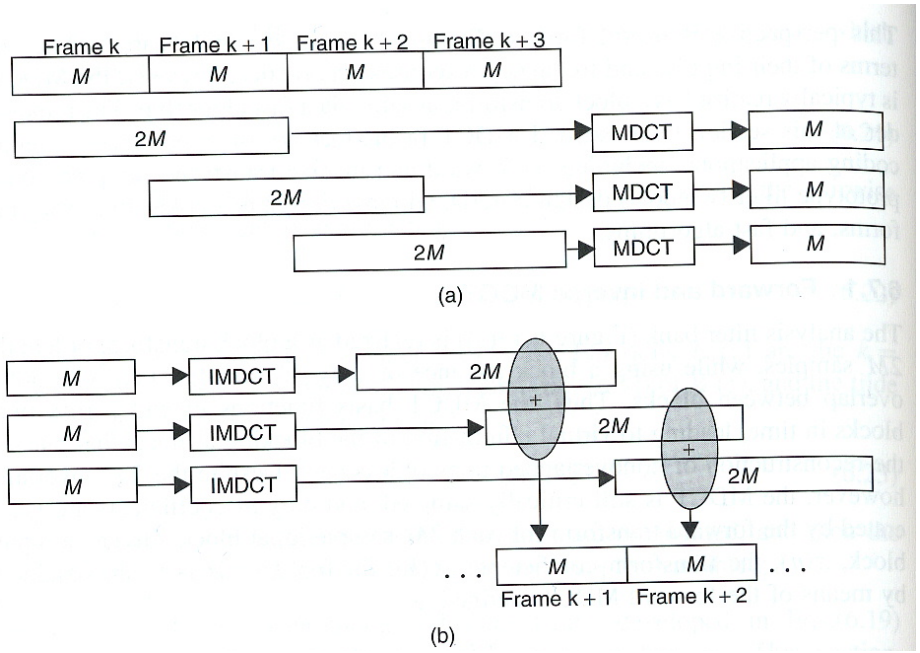


Fig: MDCT: a lapped forward transform (analysis) and inverse MDCT transform.

# MDCT window design

- $w(n)$ - FIR prototype filter (window) needs to be designed.
- Several general orthogonal and bi-orthogonal windows proposed (or designed specially for audio coding).
- Orthogonal case: PR conditions are reduced to the linear phase and Nyquist constraints on the window:

$$w(2M-1-n) = w(n)$$

$$w^2(n) + w^2(n+M) = 1$$

  for sample indices $0 \leq n \leq M-1$.
- Eqs. guarantee the orthogonal basis for the MDCT, but orthogonality is not required to satisfy PR, thus it causes less freedom for designing $w(n)$.
- Biorthogonal windows relax design of the FIR prototype filter - in effect, no longer need to use the same analysis and synthesis windows

# MDCT window design - example - Sine window

- $w(n) = sin\left[\left(n + \frac{1}{2}\right)\frac{\pi}{2M}\right]$ for $0 \leq n \leq M - 1$.
- The most popular in audio coding - MPEG1 - layer 3, MPEG-2 AAC/MPEG-4 filter banks.
- DC energy concentrated in a single transform coefficient.
- Filter bank channels achieve 24dB sidelope attenuation when sine window is used.
- Shown to be optimal in terms of coding gain for a lapped transform (quantify the factor by which MSE is reduced when using filter bank relative to using direct PCM quantization of the time-domain signal at the same rate).

# MDCT window design - example - Sine window

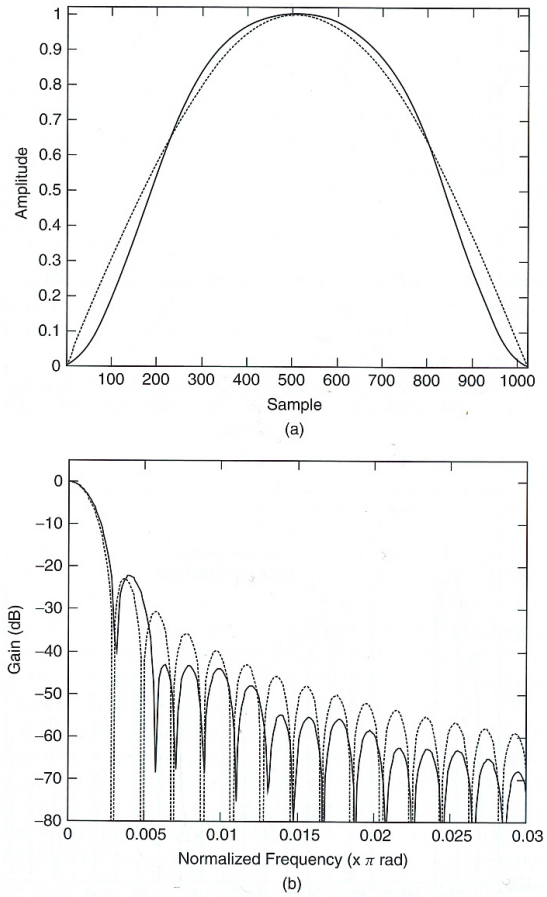


Fig: Orthogonal MDCT analysis-synthesis windows of Malwar (dashed) and Ferreira (solid): time domain and frequency domain magnitude response.

# MDCT window design - example - Sine window

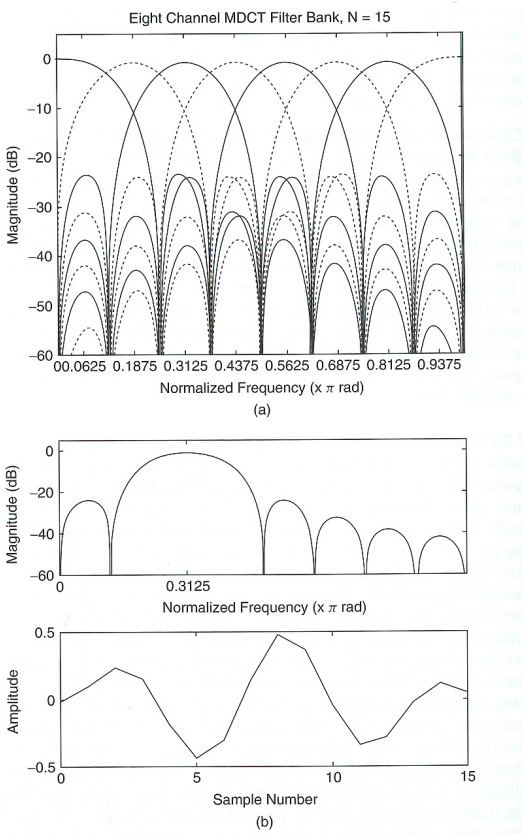


Fig: 8-channel MDCT filter bank with sine window - magnitude responses for odd and even-numbered channels. Isolated view of the magnitude freq. response and time-domain impulse response for channel 3. Assymetry is clearly visible in the channel impulse response (linear phase possibility, but overall analysis-synthesis filter bank has linear phase on all channels).

# Time-varying forms of MDCT

Particular relevance for perceptual audio codecs:

- Time-varying $\Longleftarrow$ filter bank is the best if signal-specific (e.g. tone-like and noise-like signals).
- It is very common with MDCT to change number of channels - window length to match the signal properties.
- Usually, binary classification scheme used to identify the input - either stationary or non-stationary-transient $\Longleftrightarrow$ long or short windows used.
- Long windows - maximize coding gain and achieve good channel separation.
- Short windows - localize time-domain artifacts (e.g., pre-echo).
- Very efficient - BUT complications for the codec structure.

# DFT and DCT as filter bank

Earlier, DFT and DCT often used to achieve high - resolution frequency analysis.

- FFT realization of DFT is in MPEG-1, Layer 3 - for psychoacoustic model and as well to create hybrid filter bank stages (PQMF and MDCT).
- Example - DFT: $X(k) = \frac{1}{\sqrt{2M}} \sum_{n=0}^{2M-1} x(n) W^{-nk}$, for $0 \leq k \leq 2M-1$, where $W = e^{j\pi/M}$.
- If the analysis filters have all the same length and $L = 2M$, filter bank $\sim$ block transform.
- From filter bank view point: $h_k(n) = \frac{1}{\sqrt{2M}} W^{kn}$, for $0 \leq n \leq 2M-1$, and $0 \leq k \leq M-1$.

# DFT and DCT as filter bank

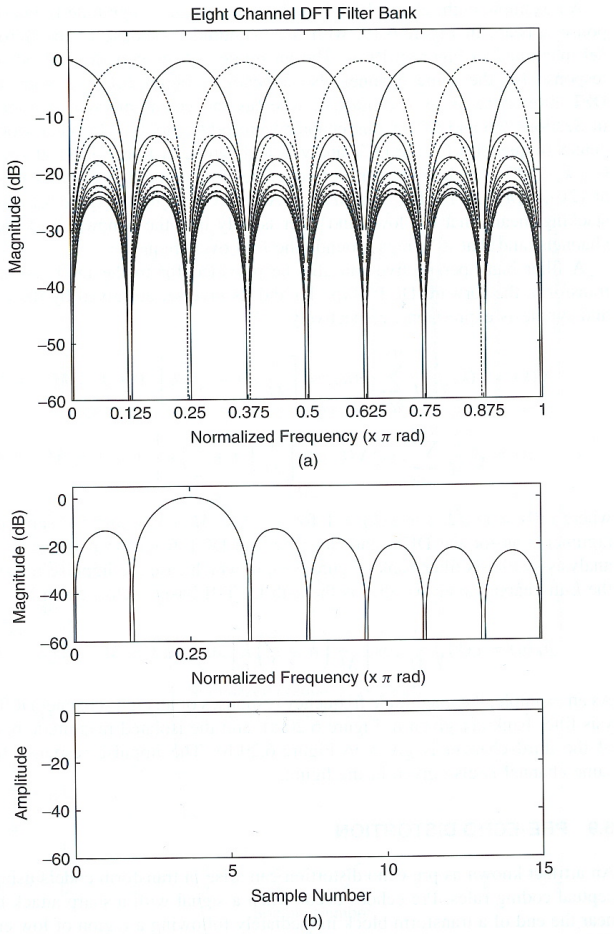


Fig: 8-band DFT- magnitude responses for odd (dashed) and even-numbered channels. Isolated view of the magnitude freq. response and time-domain impulse response for channel 3. The impulse response is complex-valued and that only its magnitude is shown.

# Pre-echo distortion

It occurs in transform coders when a signal with a sharp attack begins near the end of transform block immediately following a region of low energy (percussive instruments, such as triangle, glockenspiel, castanets). T-F uncertainty principle dictates that the inverse transform spread quantization distortion evenly in time throughout the reconstructed block (quantization is performed to satisfy the masking thresholds associate with the block average spectral estimate).

# Pre-echo distortion

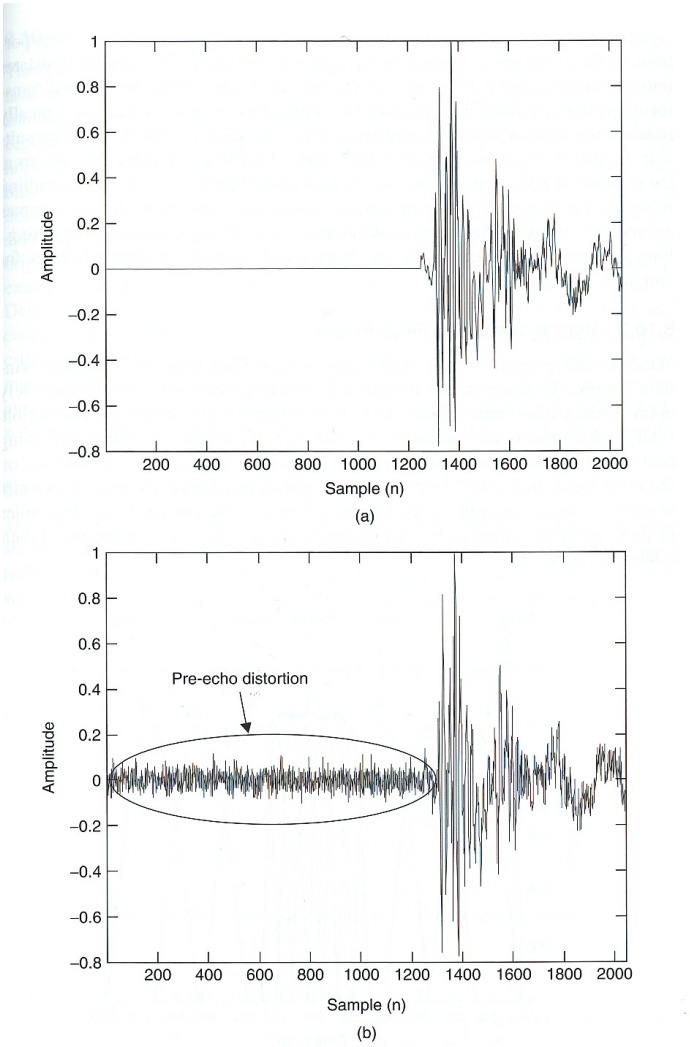


Fig: Pre-echo example - uncoded castanets and transform coded castanets (2048-point block size).

# Filter bank and lapped transforms

- If length of the filter used in M-channel sub-band filters were M samples $\Rightarrow$ linear transform of successive non-overlapping M sample frames of the input signal:

$$\begin{bmatrix} y_0(m) \\ y_1(m) \\ \dots \\ y_{M-1}(m) \end{bmatrix} = \begin{bmatrix} h_0(M-1) & \dots & h_0(1) & h_0(0) \\ h_1(M-1) & \dots & h_1(1) & h_1(0) \\ \dots & \dots & \dots & \dots \\ h_{M-1}(M-1) & \dots & h_{M-1}(1) & h_{M-1}(0) \end{bmatrix} \begin{bmatrix} x(mM-M+1) \\ \dots \\ x(mM-1) \\ x(mM) \end{bmatrix}$$

  which means $\mathbf{y}=\mathbf{H}\mathbf{x}$

- Inverse operation $\mathbf{x}=G^T\mathbf{y}$:

$$\begin{bmatrix} x(mM-M+1) \\ \dots \\ x(mM-1) \\ x(mM) \end{bmatrix} = \begin{bmatrix} g_0(M-1) & & g_1(1) & \dots & g_{M-1}(0) \\ \dots & \dots & \dots & \dots & \\ g_0(1) & & g_1(1) & \dots & \dots \\ g_0(0) & & g_1(0) & \dots & g_{M-1}(0) \end{bmatrix} \begin{bmatrix} y_0(m) \\ y_1(m) \\ \dots \\ y_{M-1}(m) \end{bmatrix}$$

# Filter bank and lapped transforms

- Non-overlapping input frames - blocking effect - therefore practical filter banks use longer analysis-synthesis filters and the transform is rather $N \times M$ or $M \times N$.

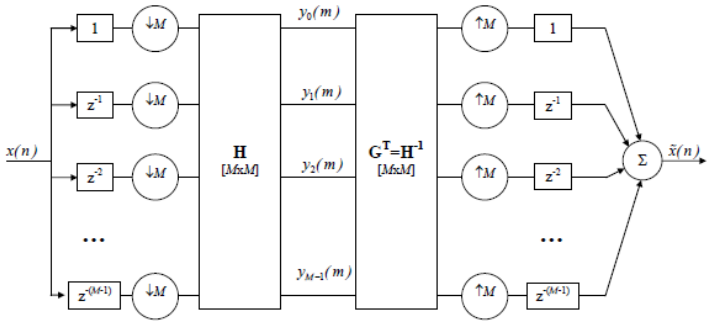


Fig: M-channel transform based implementation of a filter bank using filters of length N=M.

# Filter bank and lapped transforms

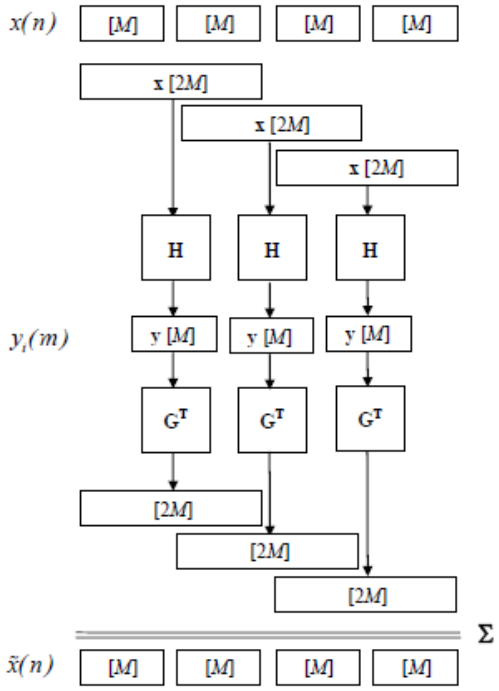


Fig: Frame based view of M-channel transform based implementation of a filter bank using filters of length N=2M.

# Filter bank and lapped transforms

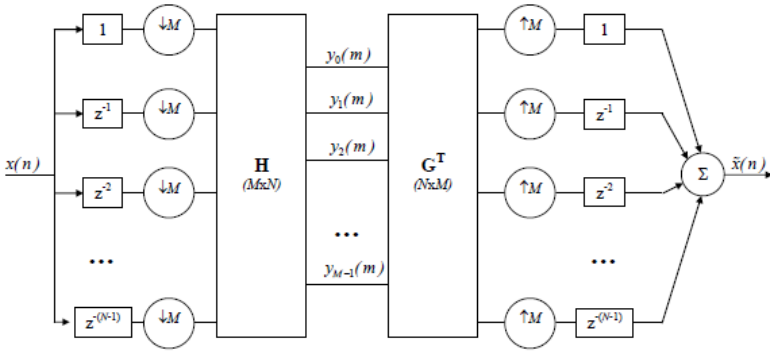


Fig: Frame based view of M-channel transform based implementation of a filter bank using filters of length N>M.

# Sub-band coding

- Exploits signal redundancy and psychoacoustic irrelevancy in the frequency domain.
- Spectrum divided into sub-bands using a band-pass filters. The output is then sampled and encoded.
- Coding gains - by efficiently quantizing decimated output sequences from PR filter banks.
- Decimated output sequences of the filter bank are normalized and quantized over short 2-5ms blocks.
- Efficient quantization relies upon psychoacoustically controlled bit-allocation rules.

# Sub-band coding

Dudley's vocoder (1939):

- No use of block processing (DFT, FFT).
- Predictability - speech waveform changes slowly (inertia of air mass in vocal tract cavities), spectral envelope changes slowly (20 Hz low-pass).
- Hearing properties - spectral resolution of hearing (wider band-pass at higher frequencies)

# Sub-band coding

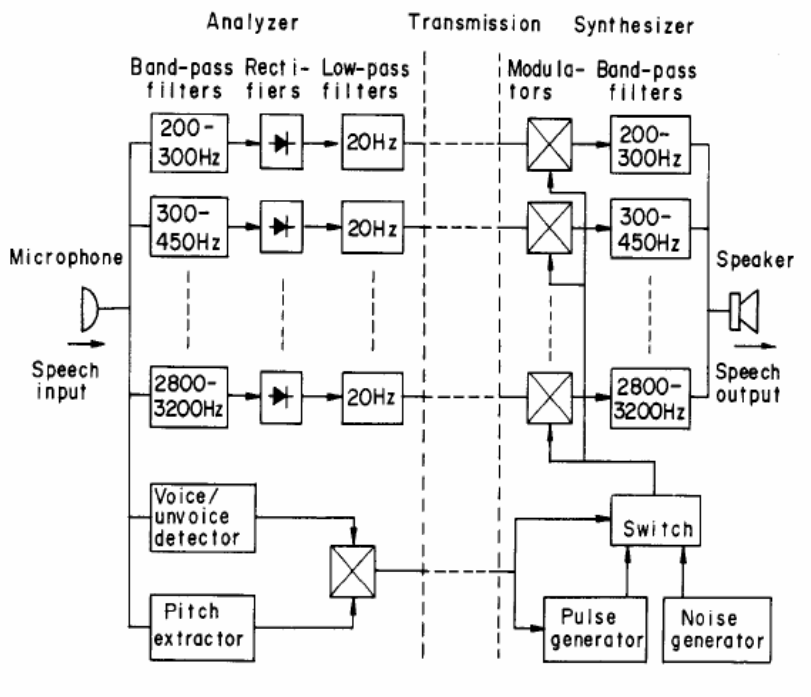


Fig: Homer Dudley (1939) - vocoder.

# Phase vocoder (Flanagan, Golden, 1966)

- Dudley's vocoder - called channel vocoder - transmits only the spectral envelope = a gross estimation of the magnitude spectrum of speech while spectral details are sent in a separate channel and modeled through a voice/unvoiced generator.
- Phase vocoder - phase information is sent together with amplitude information $\Rightarrow$ lot more information on the speech excitation.
- DFT used as an filter bank - particular case of STFT processing with special care to perfect reconstruction of the input signal $\Rightarrow$ possibility to be used for non-linear processing effects (time-scale modifications, pitch shifting, harmonization).
- DFT can be interpreted for a fixed value of $k$, as the result of passing the input signal through a sub-band filter whose frequency response $H_k(\phi)$ is centered on $\Omega_k$ and then decimated by a factor on N. Impulse frequency response is

# Phase vocoder (Flanagan, Golden, 1966)

$$h_k(n) = \{1, e^{-j\Omega_k}, e^{-j2\Omega_k}, \ldots\}.$$

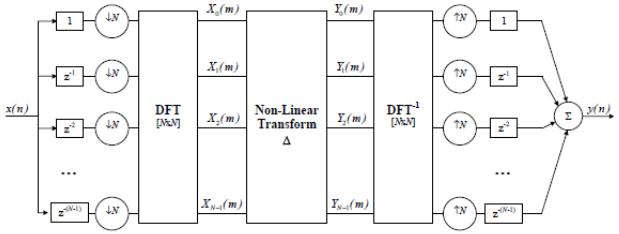


Fig: N-bin DFT-based signal processing.

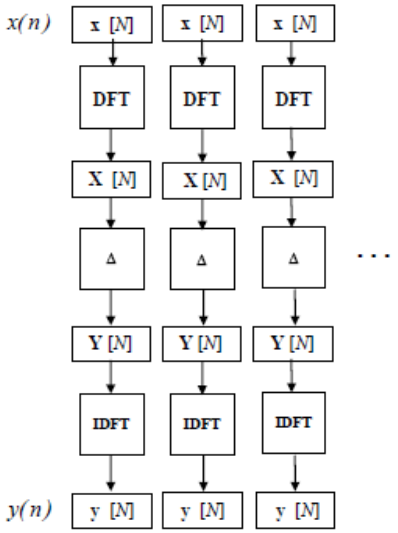


Fig: N-bin DFT-based signal processing - frame-based view.

# Phase vocoder (Flanagan, Golden, 1966)

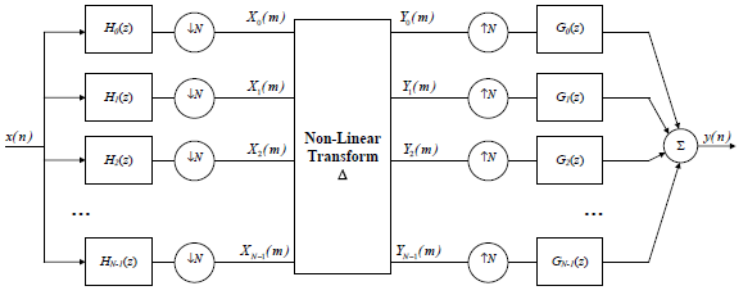


Fig: N-channel DFT-based signal processing - sub-band processing view.

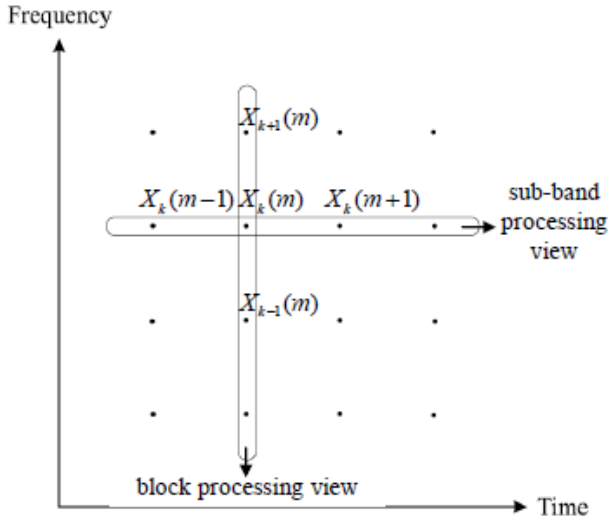


Fig: Block processing vs. sub-band processing views in time-frequency domain.

# Phase vocoder (Flanagan, Golden, 1966)

STFT-based signal processing:

- Due to non-linear processing applied to the subband signals - 3 problems may occur related to frequency selectivity, frame shift, blocking effect.
- Analysis weighting windows - DFT is not frequency selective - strong sidelobes in the frequency response of the filters - problem for phase vocoder which is assumed to have one sinusoidal components (partial) in each sub-band. Therefore weighting windows are used (Hamming, Hanning, ...).
- Overlapping analysis frames - to avoid aliasing - Nyquist theorem - sub-band signals should be sampled with sampling period smaller than N/2 samples = DFT frames should overlap with more than N/2 (N is the length of DFT filter).
- Synthesis weighting windows - synthesis based on summing overlapping output frames. In case on non-linear sub-band

# Phase vocoder (Flanagan, Golden, 1966)

processing, discontinuities appear at synthesis frame boundaries ⇒ weighting windows

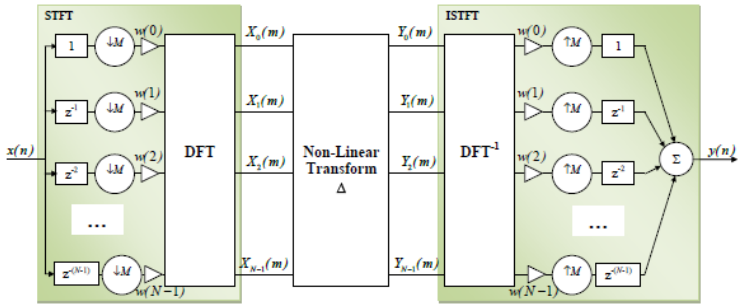


Fig: STFT based signal processing.

# Phase vocoder (Flanagan, Golden, 1966)

Time-scale modification with the phase vocoder:

- Phase vocoder - implementation of an STFT processing.
- Input is assumed to be composed of a sum of sinusoidal (not necessarilly harmonic) terms named partials:

$$x(n) = \sum_{i=1}^{P} A_i cos(n\varphi_i + \phi_i)$$

  $\varphi_i = \omega_i / F_s$ - angular frequency [rad/sample]. $n\varphi_i + \phi_i$ gives instantaneous phase of the partial $i$ at sample $n$.

- The output of each DFT is assumed to be influenced by a single partial $\Rightarrow$ long input frames.
- Then, the output of each DFT channel is a complex exponential function over $m$. If we assume simple imaginary exponential:

$$x(n) = Ae^{jn\varphi+\phi}$$

# Phase vocoder (Flanagan, Golden, 1966)

then $X_k(m)$ is:

$$X_k(m) = \sum_{n=0}^{N-1} A e^{j(n+mM)\varphi+\phi} w(n) e^{-jn\Omega_k}$$

$$= H_k(\varphi) e^{j(mM\varphi+\phi)}$$

$H_k(\varphi)$ - frequency response of the k-th analysis sub-band filter for frequency $\varphi$ - does not depend on $m$.

- Thus output of each DFT channel is an imaginary exponential function depending only on the frequency of the partial $\varphi$ but not on the central frequency $\Omega_k$ of the DFT bin.
- Therefore:

$$\measuredangle X_k(m+1) = \measuredangle X_k(m) + M\varphi$$

$$|X_k(m+1)| = |X_k(m)|$$

# Phase vocoder (Flanagan, Golden, 1966)

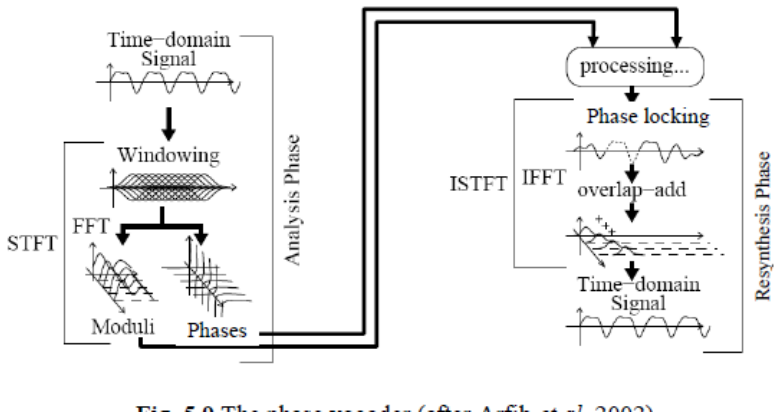


Fig: phase vocoder.

Horizontal phase locking to modify the duration of the input signal:

- Analysis frame-shift $M_a$ is lower (time expansion) or higher (time compression) than the synthesis frame shift $M_s$, and

$$\measuredangle Y_k(m+1) = \measuredangle Y_k(m) + M_s \varphi_k(m)$$

$$|Y_k(m)| = |X_k(m)|$$

Horizontal and vertical phase locking:

- Phasiness - characteristic coloration of the signal (speech

# Phase vocoder (Flanagan, Golden, 1966)

sounds like the speaker is much further from mic than original recording. It can be minimized by ensuring also phase coherency across channels in the given frame (not only phase consistency within each DFT channel over time).

- $\Rightarrow$ Phase locked vocoder.

# MUSICAM

Masking Pattern Adapted Universal Subband Integrated Coding and Multiplexing (MUSICAM, 1990):

- formed basis for MPEG-1 and MPEG-2 audio layers.
- Uses uniform bandwidth 32-band PQMF, delay around 10.66ms, thus no critical band processing.
- Enhanced psychoacoustic analysis (1024 - FFT in parallel with PQMF.

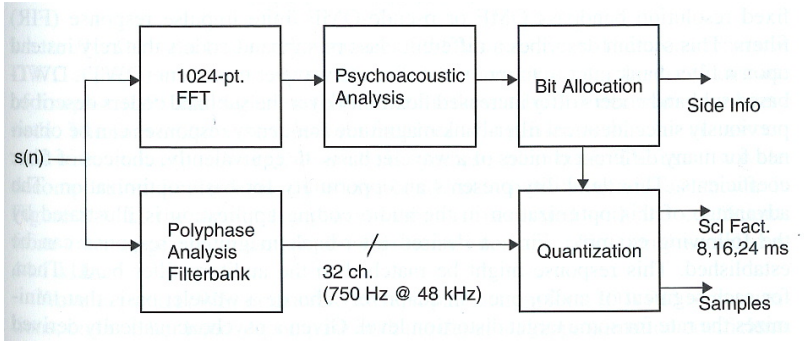


Fig: MUSICAM encoder.

# Transform coding

- In many ways, transform and sub-band coder categories overlap and sometime it is hard to categorize.
- Typically, transform coders perform high-resolution frequency analysis; sub-band coders rely on a coarse division of the frequency spectrum.
- TC uses unitary transforms (DFT, DCT, etc.) for the time-frequency analysis.
- Therefore, usually buferring at the beginning to the blocks.

# Adaptive Transform Coding (ATC) for speech

- Uses DCT (short-term spectrum), adaptive quantization and bit assignment rules.
- Coarse description of the spectrum for each frame is transmitted (as side information).
- An estimate of the short-spectrum is formed using linear interpolation in the log-domain and used to determine optimal bit assignment.
- For 16-32kbps, it outperforms the log-PCM by 17-23dB (SNR).

# Adaptive Transform Coding (ATC) for speech

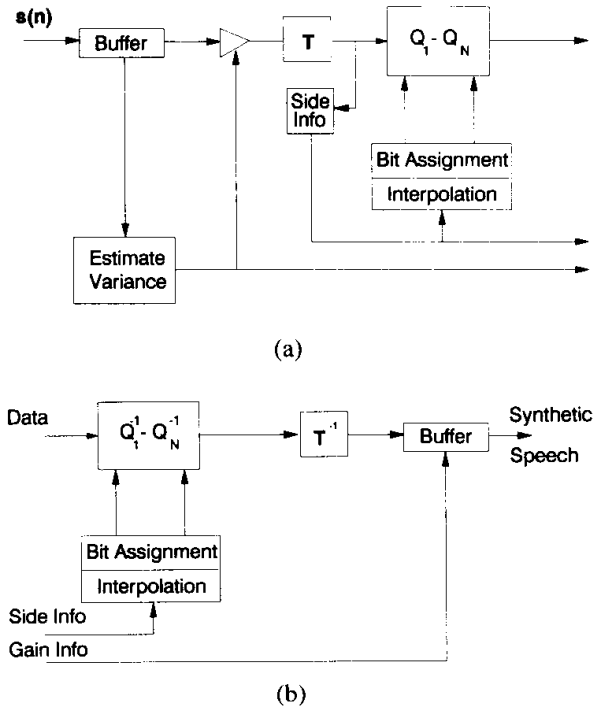


Fig: ATC coder and decoder.

# Perceptual Transform Coder (PXFM)

- Estimates amount of quantization noise that can be inaudibly injected into each transform domain subband using PE estimate.
- Windowed overlapping (1/16) segments are transformed using 2048-point FFT, JND thresholds estimated for each critical band.
- Iterative 128 sub-band quantizer is adapted using iterative quantization loop - to satisfy JND thresholds until the fixed bit-rate is achieved. The quantization and bit-packing is performed.

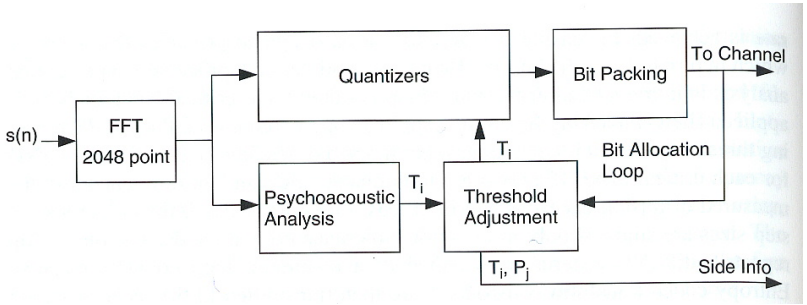


Fig: PXFM encoder.