Perceptual Audio Coding

Henrique Malvar
Managing Director, Redmond Lab
Microsoft Research

UW Lecture – December 6, 2007

Contents

- Motivation
- "Source coding": good for speech
- "Sink coding": Auditory Masking
- Block & Lapped Transforms
- Audio compression
- Examples

Microsoft Research 2 University of Washington

Contents

- Motivation
- “Source coding”: good for speech
- “Sink coding”: Auditory Masking
- Block & Lapped Transforms
- Audio compression
- Examples

Microsoft Research 3 UNIVERSITY OF WASHINGTON

Many applications need digital audio

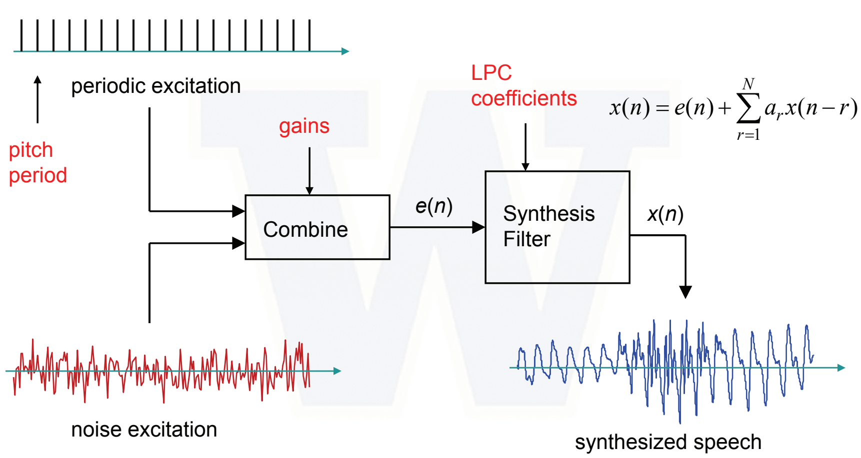
- Communication
  - Digital TV, Telephony (VoIP) & teleconferencing
  - Voice mail, voice annotations on e-mail, voice recording
- Business
  - Internet call centers
  - Multimedia presentations
- Entertainment
  - 150 songs on standard CD
  - thousands of songs on portable music players
  - Internet / Satellite radio, HD Radio
  - Games, DVD Movies

Microsoft Research 4 UNIVERSITY OF WASHINGTON

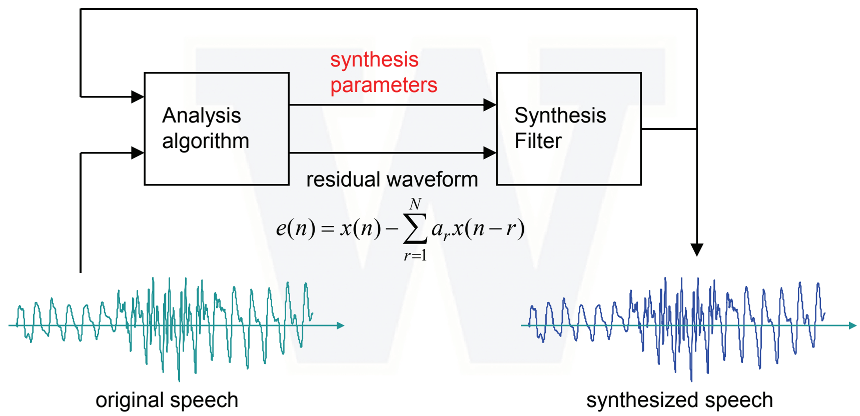
Contents

- Motivation
- "Source coding": good for speech
- "Sink coding": Auditory Masking
- Block & Lapped Transforms
- Audio compression
- Examples

Linear Predictive Coding (LPC)

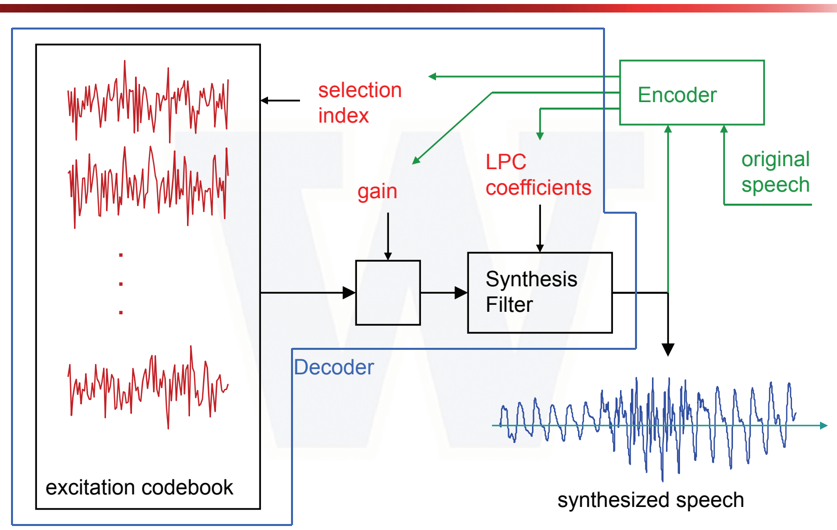


LPC basics – analysis/synthesis

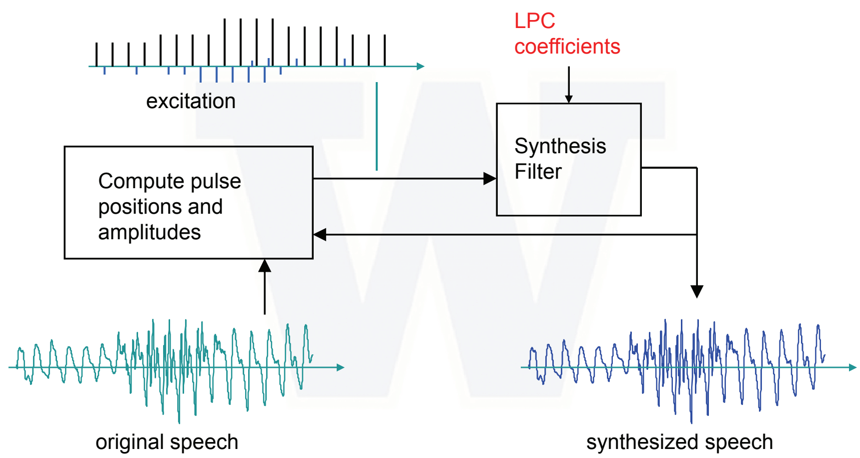


$$e(n) = x(n) - \sum_{r=1}^{N} a_r x(n-r)$$

LPC variant - CELP



LPC variant - multipulse

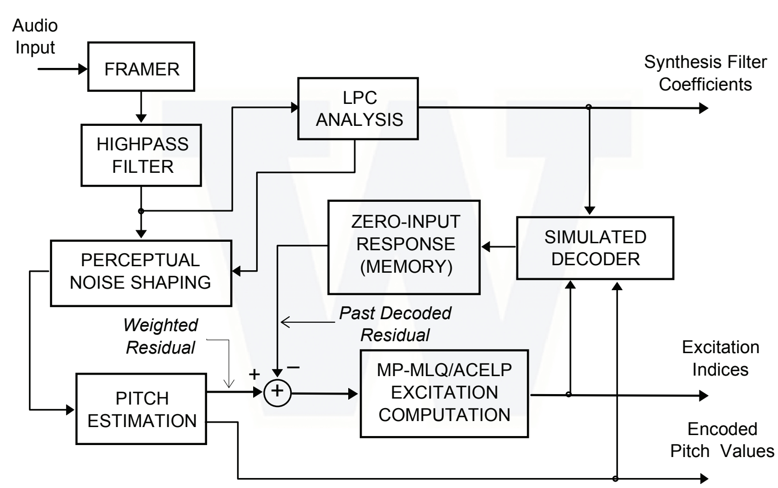


Microsoft Research

9

UNIVERSITY OF WASHINGTON

G.723.1 architecture



Microsoft Research

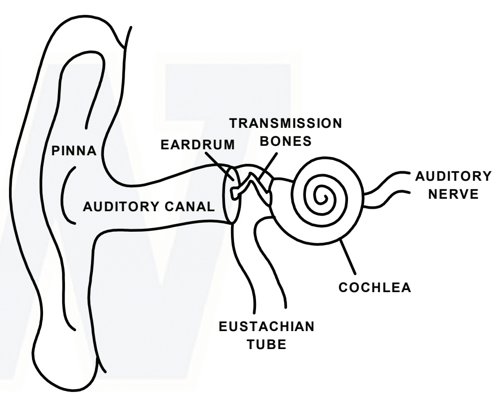
10

UNIVERSITY OF WASHINGTON

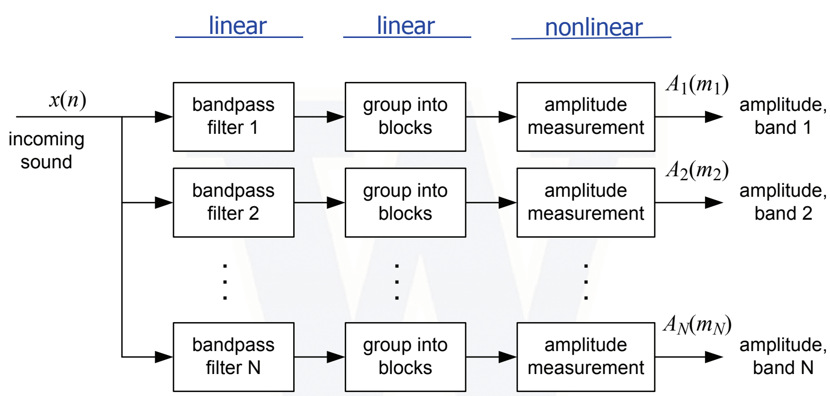
Contents

Physiology of the ear

- Automatic gain control
  - muscles around transmission bones
- Directivity
  - pinna
- Boost of middle frequencies
  - auditory canal
- Nonlinear processing
  - auditory nerve
- Filter bank separation
  - cochlea
- Thousands of "microphones"
  - hair cells in cochlea



Filter bank model



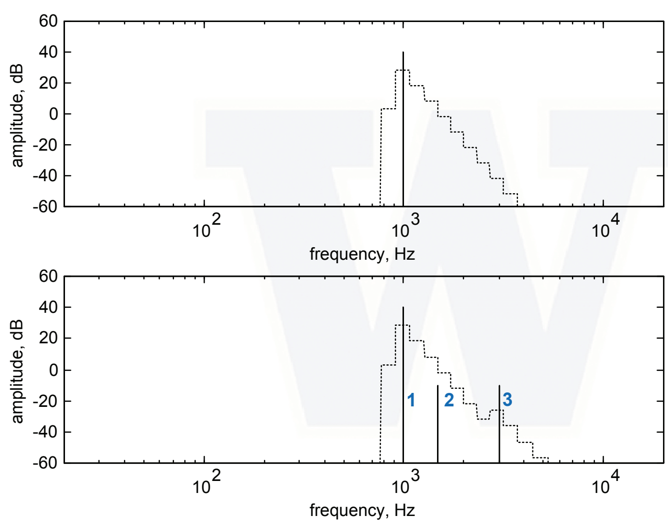
- Explains frequency-domain masking

Microsoft Research

13

UNIVERSITY OF WASHINGTON

Frequency-domain masking



- tone 1
- tone 2
- tone 3
- tone 1+3
- tone 1+2+3

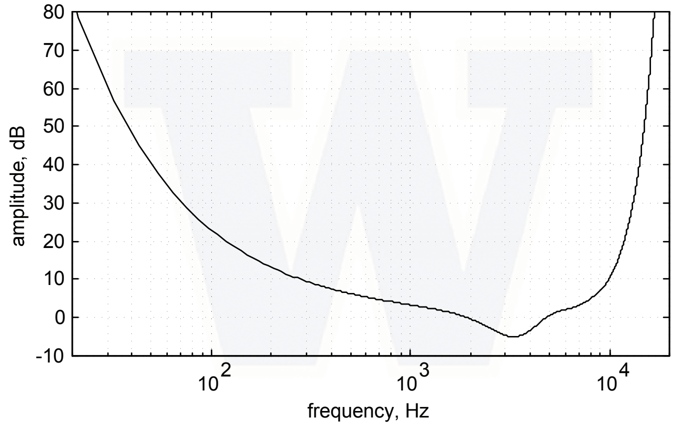
Microsoft Research

14

UNIVERSITY OF WASHINGTON

Absolute threshold of hearing

- Fletcher-Munson curves

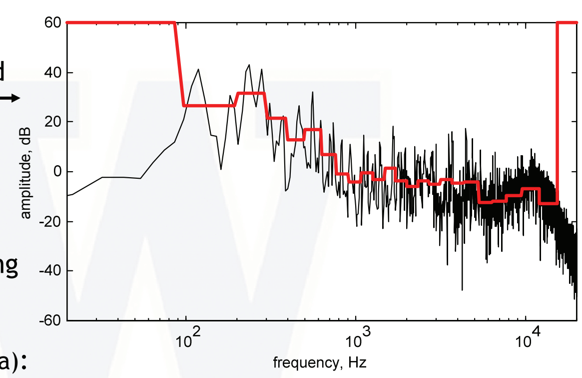


- Basis for loudness correction in audio amplifiers

15

Example of masking

- Typical spectrum & masking threshold →
- Original sound:
- Sound after removing components below the threshold (1/3 to 1/2 of the data):

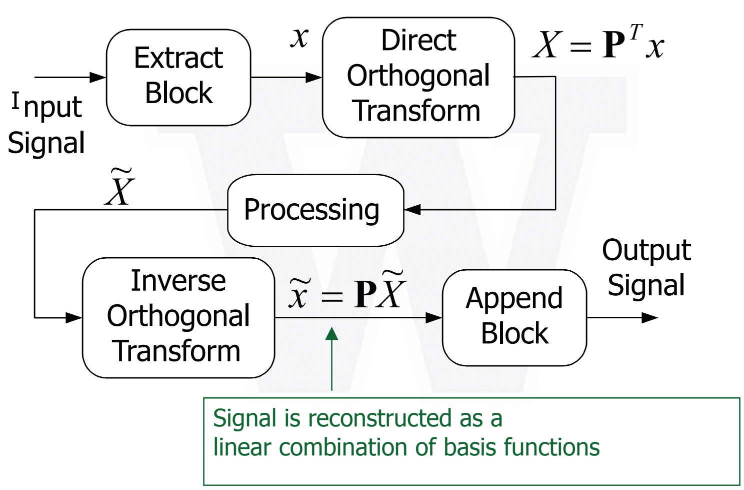


16

Contents

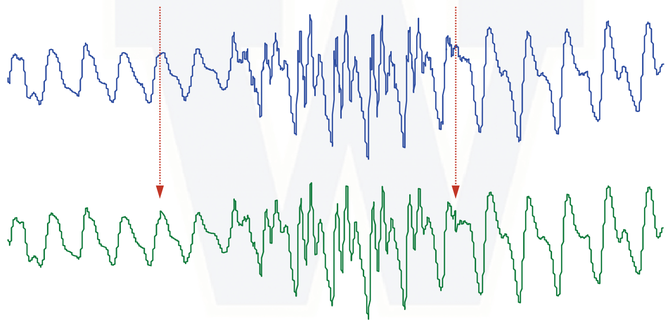
- Motivation
- "Source coding": good for speech
- "Sink coding": Auditory Masking
- **Block & Lapped Transforms**
- Audio compression
- Examples

Block signal processing



Block processing: good and bad

- Pro: allows adaptability



- Con: blocking artifacts

Why transforms?

- More efficient signal representation
  - Frequency domain
  - Basis functions ~ "typical" signal components
- Faster processing
  - Filtering, compression
- Orthogonality
  - Energy preservation
  - Robustness to quantization

Compactness of representation

- Maximum energy concentration in as few coefficients as possible
- For stationary random signals, the optimal basis is the Karhunen-Loève transform:

$$\lambda_i p_i = R_{xx} p_i, \quad \mathbf{P}^{\mathbf{T}}\mathbf{P} = \mathbf{I}$$

- Basis functions are the columns of $\mathbf{P}$
- Minimum geometric mean of transform coefficient variances

Microsoft Research

21

UNIVERSITY OF WASHINGTON

Sub-optimal transforms

- KLT problems:
  - Signal dependency
  - $\mathbf{P}$ not factorable into sparse components
- Sinusoidal transforms:
  - Asymptotically optimal for large blocks
  - Frequency component interpretation
  - Sparse factors - e.g. FFT

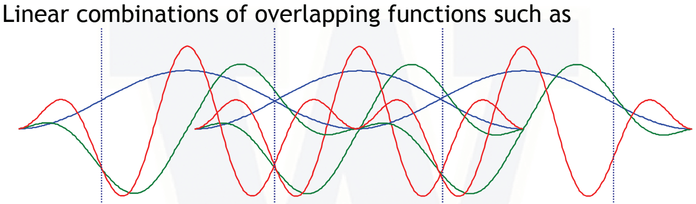
Microsoft Research

22

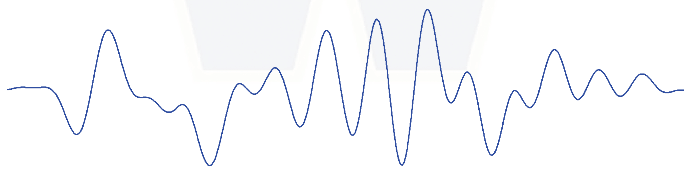
UNIVERSITY OF WASHINGTON

Lapped transforms

- Basis functions have tails beyond block boundaries
  - Linear combinations of overlapping functions such as



  - generate smooth signals, without blocking artifacts



Microsoft Research 23 UNIVERSITY OF WASHINGTON

Modulated lapped transforms

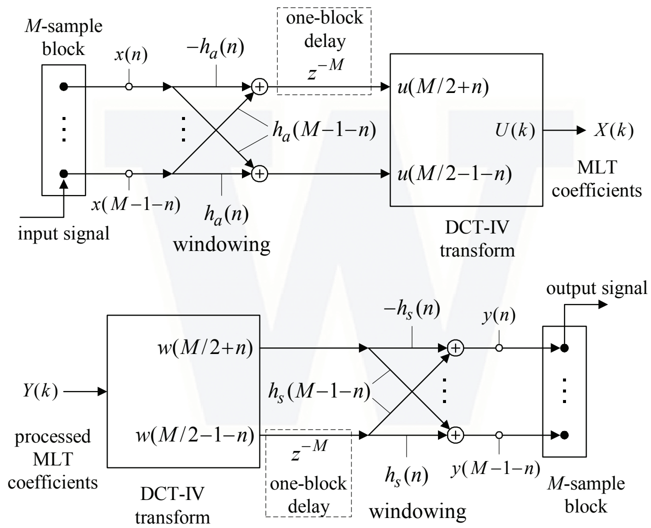
- Basis functions = cosines modulating the same low-pass (window) prototype $h(n)$:

$$p_k(n) = h(n)\sqrt{\frac{2}{M}}\cos\left[\left(n+\frac{M+1}{2}\right)\left(k+\frac{1}{2}\right)\frac{\pi}{M}\right]$$

- Can be computed from the DCT or FFT
- Projection $X = \mathbf{P}^T x$ can be computed in $O(\log_2 M)$ operations per input point

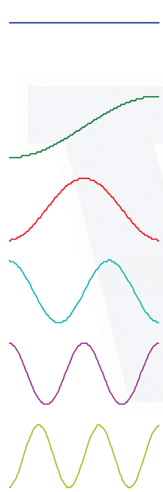
Microsoft Research 24 UNIVERSITY OF WASHINGTON

Fast MLT computation

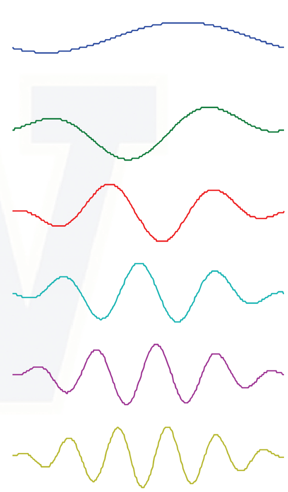


Basis functions

DCT:



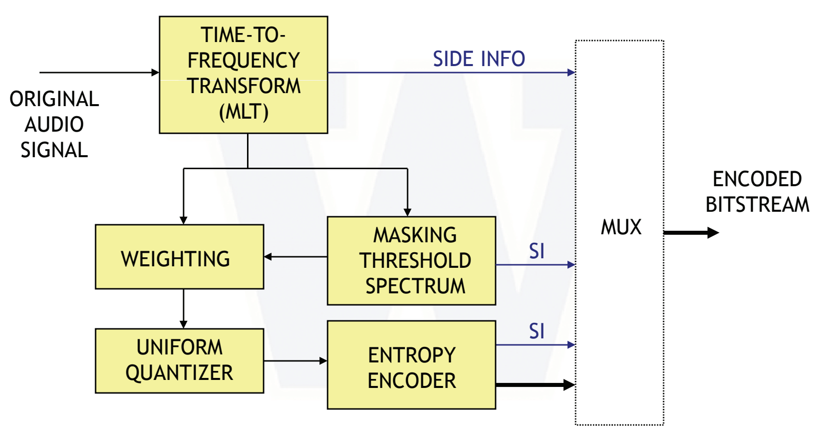
MLT:



Contents

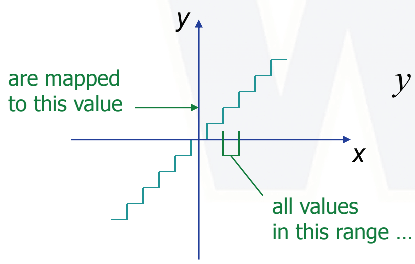
- Motivation
- "Source coding": good for speech
- "Sink coding": Auditory Masking
- Block & Lapped Transforms
- Audio compression
- Examples

Basic architecture



Quantization of transform coefficients

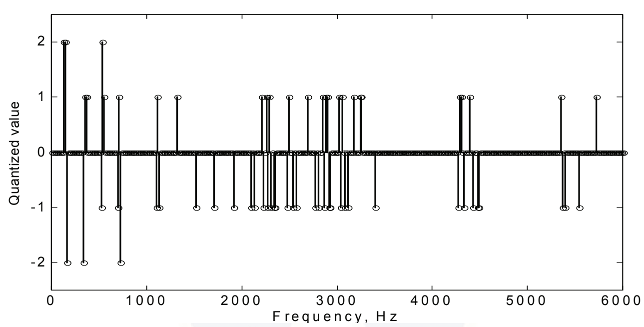
- Quantization = rounding to nearest integer.
- Small range of integer values = fewer bits needed to represent data
- Step size T controls range of integer values



$$y = T\,\mathrm{int}(x / T)$$

Encoding of quantized coefficients

- Typical plot of quantized transform coefficients



- Run-length + entropy coding

Basic entropy coding

- Huffman coding: less frequent values have longer codewords
- More efficient if groups of values are assembled in a vector before coding

Value	Codeword
−7	'1010101010001'
−6	'10101010101'
−5	'101010100'
−4	'10101011'
−3	'101011'
−2	'1011'
−1	'01'
0	'11'
+1	'00'
+2	'100'
+3	'10100'
+4	'1010100'
+5	'1010101011'
+6	'101010101001'
+7	'1010101010000'

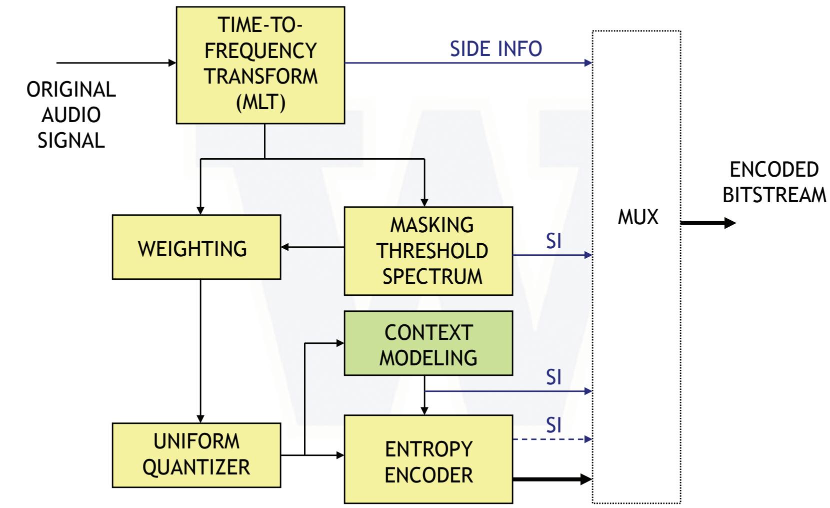
Microsoft Research 31 UNIVERSITY OF WASHINGTON

Side information & more about EC

- Side info: model of frequency spectrum
  - e.g. averages over subbands
- Quantized spectral model determines weighting
  - masking level used to scale coefficients
- Backward adaptation reduces need for SI
- Run-length + Vector Huffman works
  - Context-based AC can be better
  - Room for better context models via machine learning?

Microsoft Research 32 UNIVERSITY OF WASHINGTON

Improved architecture



Examples of context modeling

- For strongly voiced segments, spectral energies may be well predicted by a "Linear Prediction" model, similar to those used in VoIP coders.
- For strongly periodic components, spectral energies may be predicted by a pitch model.
- For noisy segments, a noise-only model may allow for very coarse quantization → lower data rate.

Other aspects & directions

- Stereo coding
  - (L+R)/2 & L-R coding, expandable to multichannel
  - Intensity + balance coding
  - Mode switching – extra work for encoder only
- Lossless coding
  - Easily achievable via integer transforms
  - exactly reversible via integer arithmetic
  - example: lifting-based MLT (see Refs)
- Using complex subband decompositions (MCLT)
  - Potential for more sophisticated auditory models
  - Efficient encoding is an open problem

Audio coding standards

ISO/IEC	MPEG-1 Layer III (MP3) · MPEG-1 Layer II · MPEG-1 Layer I · **AAC** · HE-AAC · HE-AAC v2
ITU-T	G.711 · G.722 · G.722.1 · G.722.2 · G.723 · G.723.1 · G.726 · G.728 · G.729 · G.729.1 · G.729[a]
Others	AC3 · AMR · Apple Lossless · ATRAC · FLAC · iLBC · Monkey's Audio · μ-law · Musepack · Nellymoser · OptimFROG · RealAudio · RTAudio · SHN · Speex · Vorbis · WavPack · WMA · TAK

From http://en.wikipedia.org/wiki/Advanced_Audio_Coding

Contents

- Motivation
- "Source coding": good for speech
- "Sink coding": Auditory Masking
- Block & Lapped Transforms
- Audio compression
- Examples

Microsoft Research

UNIVERSITY OF WASHINGTON

WMA examples:

- Original clip (~1,400 kbps)　　64 kbps (MP3)　　64 kbps (WMA)
- Original clip　　WMA @ 32 kbps (Internet radio)
- More examples at

http://www.microsoft.com/windows/windowsmedia/demos/audio_quality_demos.aspx

Microsoft Research

UNIVERSITY OF WASHINGTON

References

- S. Shlien, “The modulated lapped transform, its time-varying forms, and its applications to audio coding standards,” *IEEE Trans. Speech and Audio Processing*, vol. 5, pp. 359-366, July 1997.
- H. S. Malvar, “Fast Algorithms for Orthogonal and Biorthogonal Modulated Lapped Transforms,” *IEEE Symposium Advances Digital Filtering and Signal Processing*, Victoria, Canada, pp. 159–163, June 1998.
- H. S. Malvar, “Enhancing the performance of subband audio coders for speech signals,” *IEEE International Symposium on Circuits and Systems,* Monterey, CA, vol.5, pp. 98–101, June 1998.
- H. S. Malvar, “A modulated complex lapped transform and its applications to audio processing,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, AZ, pp. 1421–1424, March 1999.
- T. Painter and A. Spanias, “Perceptual coding of digital audio,” *Proc. IEEE*, vol. 88, pp. 451–513, Apr. 2000. Available at http://www.eas.asu.edu/~spanias/papers.html
- H. S. Malvar, “Auditory Masking in Audio Compression,” chapter in *Audio Anecdotes*, K. Greenebaum, Ed., A. K. Peters Ltd., 2004.
- J. Li, “Reversible FFT and MDCT via matrix lifting,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, pp. IV-173–176, May 2004.
- H. S. Malvar, “Adaptive run-length/Golomb-Rice encoding of quantized generalized Gaussian sources with unknown statistics,” *IEEE Data Compression Conference*, Snowbird, UT, March 2006.
- http://en.wikipedia.org/wiki/Advanced_Audio_Coding

Microsoft Research

UNIVERSITY OF WASHINGTON