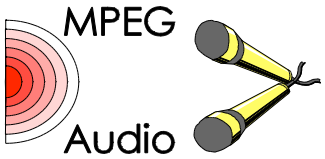


Using MPEG-7 Audio Low Level Descriptor Scalability: A Guided Tour

Jürgen Herre, Eric Allamanche
Fraunhofer Institut for Integrated Circuits (FhG-IIS)
Erlangen, Germany

Overview

- The need for scalable metadata
- MPEG-7 Audio scalability
  - Scalability over frequency
  - Scalability over time -> scalable series

- The ScalableSeries
  - Concept
  - Application examples
  - Limitations

- Conclusions

Fraunhofer Institut
Integrierte Schaltungen

The Need for Metadata Scalability

Rich / precise descriptions

Compact/coarse descriptions

- Offer highest retrieval performance
- Occupy large amount of data
- High computational matching effort

- Reduced retrieval performance
- Smaller description database
- Manageable matching effort

- There is no fixed ‘one size fits all applications’ trade-off!
- MPEG-7 Audio format needs to serve all conceivable applications

=> Concept of “Scalability”

Fraunhofer Institut
Integrierte Schaltungen

Low Level Descriptor Overview

Basic	• Instantaneous waveform, power, silence
Basic Spectral	• Power spectrum, spectral centroid, spectral spread, spectral flatness …
Signal Parameters	• Fundamental frequency, harmonicity
Spectral Basis Representations	• Used primarily for sound recognition, projections into low-dimensional space
Timbral Temporal	• Log attack time, temporal centroid of a monophonic sound
Timbral Spectral	• Features specific to the harmonic portions of signals (harmonic spectral centroid, spectral deviation, spectral spread, …) etc.

Fraunhofer Institut
Integrierte Schaltungen

Metadata Scalability

General Concept

- Many descriptive elements are *scalable* in one or more dimensions
  - Derive reduced-rate (compact) precision description from original rich version [may happen on-the-fly, e.g. streaming apps]
  - Result of "downscaling" is still compatible to original version (i.e. can be compared in a meaningful way)
  - *Hierarchical* scalability concept, i.e. richer representations are supersets of smaller ones

(see also "MPEG-4 Scalability", i.e. parts of A/V bitstream can be dropped => lower quality reconstruction of signal)

Fraunhofer Institut
Integrierte Schaltungen

MPEG-7 Audio Scalability (2)

Spectral Scalability

- Some LLDs (e.g. `AudioSpectrumEnvelope`) have *spectral resolution / coverage* as scalability dimension
  - Resolutions are powers of 2 (1/16 … 8 oct.)
  - Frequency bands anchored around 1kHz

  => Commensurate frequency band definitions
  => All conceivable descriptions can be compared to each other
- Easy downscaling via simple additive combination of adjacent band energies

Fraunhofer Institut
Integrierte Schaltungen

MPEG-7 Audio Scalability Concepts (3)

Temporal Scalability

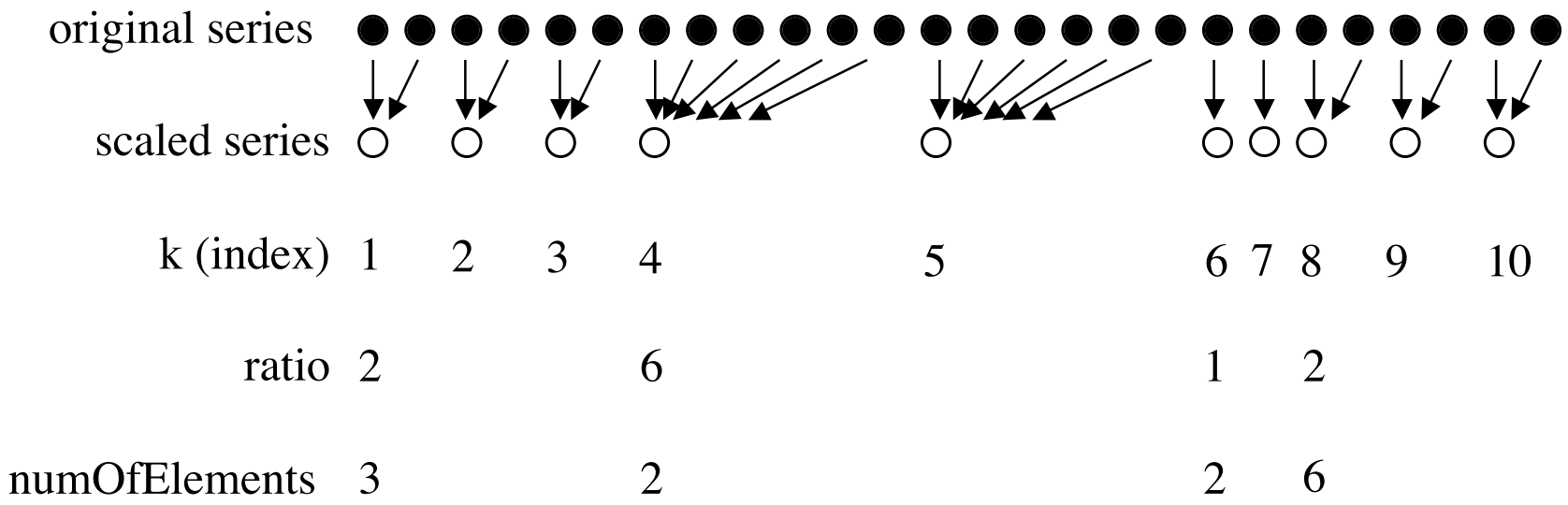
- Many LLDs describe time-varying signal properties sampled time series
- How fine sampling is needed ?
  (depends on individual application
   cannot be fixed at time of definition)
- Sampling rate determines data rate

Approach

- Define *one* generic construct for sampled time series which enables a scalable representation [Alain de Cheveigné]:

   Concept of `ScalableSeries`

Scalable Series



General Idea: Variable decimation of sampled data (factor `ratio`)

Scalable Series (2)

Concept of `ScalableSeries`

- Allow downsampling of sampled descriptor data in time
- Decimation may happen in several steps, yet results are independent of intermediate steps
- Provides summarization of raw data
  - `Mean, Variance`
  - `Min, Max, First, Last, Random`
- Provides additional hints
  - `Weigth`

Fraunhofer Institut
Integrierte Schaltungen

Scalable Series (3)

Data types

- Available for
  - scalar data types (e.g. AudioSpectrumCentroid) `SeriesofScalar`
  - vector data types (e.g. subband-based data such as AudioSpectrumEnvelope) `SeriesofVector`
- Subtypes: Decimation by powers of 2
  - `SeriesofScalarBinary`, `SeriesofVectorBinary`
  - Strictly hierarchical
  - Allows to retain some statistics of what has been lost by progressive downsampling (“`scalewise variance`”)

Application Example: Scalable Audio Fingerprint

LLD/Feature

- `AudioSignature/AudioSpectrumFlatness`, data type `SeriesofVectorBinary`

Offers different “operating points”:

Time

- Duration / Time Coverage

Granularity

- Temporal Granularity / Resolution: 60ms ... 1s ... 4s

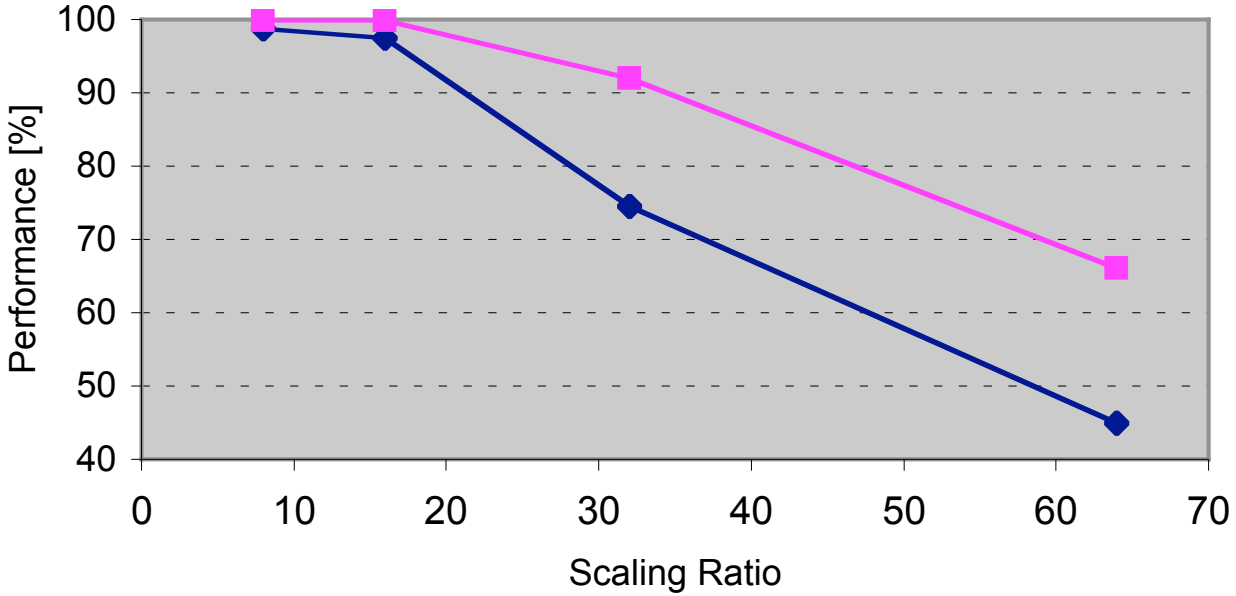
Richness

- Spectral Information Amount (number of frequency bands): 4 ... 16 ... 24

- Datarate of 2Byte/s ... 32 ... 800Byte/s, Fingerprints can be “transcoded”

Fraunhofer Institut
Integrierte Schaltungen

Application Example (2)



- Performance of an audio fingerprinting application (`ASF` LLD) as a function of temporal resolution of description (`ratio`)
- Varying scaling ratio between 8 and 64

Fraunhofer Institut
Integrierte Schaltungen

Application Example: Search in Large Databases

Fast Preselection Search

- Varying the spectral coverage
- Database: 15,000;
- Test: 1,000 Items; Length 20s ; Top2%

No. of bands	4	8	16(def.)
Speed up factor	4	2	1
MP3@96kbps	100.0%	100.0%	100.0%
Microphone	100.0%	99.9%	99.9%
Resampling	99.4%	100.0%	100.0%
Equalizer	97.3%	99.9%	100.0%
DynComp	100.0%	100.0%	100.0%

Fraunhofer Institut
Integrierte Schaltungen

Application Example: Search in Large Databases

Fast Preselection Search

- Decreasing the temporal resolution
- Database: 15,000;
- Test: 1,000 Items; Length 20s ; Top2%

Grouping	32(def. 0.96s)	128 (3.84s)	256 (7.68s)
Speed up factor	1	16	64
MP3@96kbps	100.0%	100.0%	99.1%
Microphone	99.9%	99.6%	98.7%
Resampling	100.0%	100.0%	98.4%
Equalizer	100.0%	100.0%	96.9%
DynComp	100.0%	99.9%	99.5%

Fraunhofer Institut
Integrierte Schaltungen

Scalable Series: Limitations

- Most important data field: `mean`
- Can be modeled as a Moving Average (MA) filter with subsequent decimation stage
- Decimation ratio *N* is equal to the number of grouped data points

Advantages

- Simplicity; computation in several (intermediate) steps leads to same result

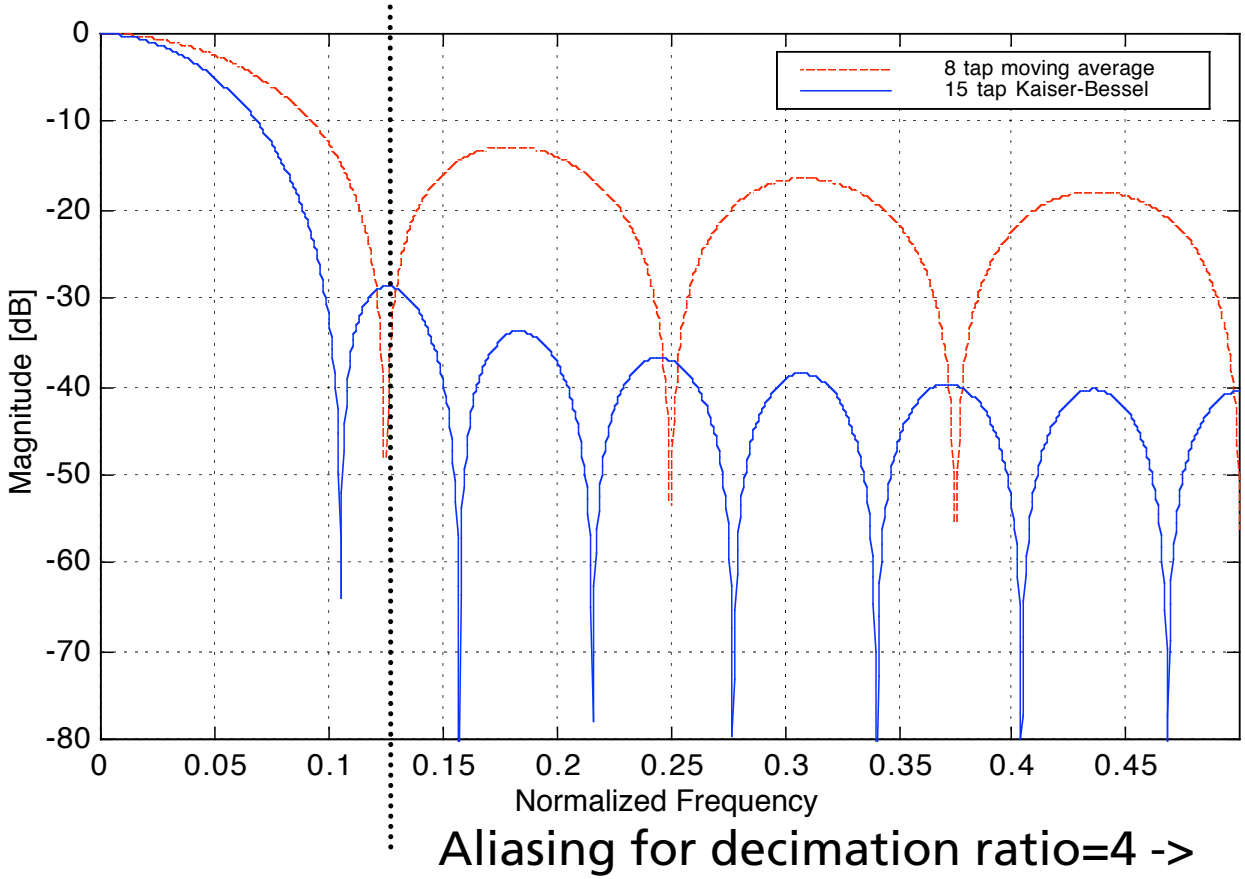
Limitations

- Poor filter attenuation above cut-off frequency => aliasing after decimation!

Question

- Performance penalty for applications due to aliasing components?

Scalable Series: Filter Frequency Response



Aliasing for decimation ratio=4 ->

Fraunhofer Institut
Integrierte Schaltungen

A Test Case: Audio Fingerprinting For Cell Phone

Test Database	• 440 audio excerpts recorded over GSM cell-phones (10-20s duration, highly distorted)
Reference items	• 15,000 fingerprints extracted from high quality audio material (CD)
Feature Setup	• ASE, 250Hz - 4000Hz; 2,1,1/2,1/4 octaves ⇒ total of 2,4,8,16 bands • Decimation by factor of 4 for recognition • Comparing application performance: – Standard ScalableSeries Decimation (MA) – Decimation after 15 taps Kaiser-Bessel LP

Results

Resolution [octaves]	2	1	1/2	1/4
Moving Average	95.07%	97.98%	98.20%	97.98%
Kaiser-Bessel	96.63%	98.65%	98.43%	98.43%

- Some performance penalty, may not be significant for many applications

Remedy: Use lower `ScalableSeries` decimation factor …

Conclusions

- Scalability is essential for a universal metadata format - allows flexible trade-off between computational efficiency/ compactness and precision
- MPEG-7 Audio includes several types of scalability
  - Frequency scalability
  - Time resolution (via ScalableSeries)
- ScalableSeries is simple & efficient way to decimate data over time (small penalty)
- Both scalability types have clearly proven their usefulness in applications …