Parametric Audio Processing

Gerald Schuller

Ilmenau University of Technology
and Fraunhofer Institute for Digital Media Technology (IDMT)

May 16, 2017
1 Introduction

2 Filter Banks

3 Musical Instrument Coder

4 Instrument Separation

5 Conclusions
Introduction

- Audio coding uses more lower bit rates, driven for instance by mobile devices,
- at the same time more functionality because of higher available processing power
- Application are surround sound, recognition of songs, meaningful modifications (multichannel from stereo or mono, movement or modification of instruments)
- Similarly in image and video processing, models are useful for compression or separation or artifacts reduction
- For instance compressive sensing uses models with whom the expected signal can be represented with few non-zero parameters
Introduction

- Goal: Audio should sound "natural", without obvious artifacts, but not necessarily exactly like the original.

- Similar for visual 3-D objects: widespread use for instance in games, games have recently about as much revenue as movies, or in animated movies.

- Similar goal: the models should look natural, without obvious artifacts, but not necessarily like the original.

- This leaves possibilities for saving bits and allow natural sounding or looking modifications, or to analyse or separate objects in this "model space".
Goal: generate natural sounding audio object with as few parameters as possible

Approach: use suitable models

In this presentation I will go to more details for filter banks for audio coding and give shorter examples for other applications
First Example: Filter Banks for Audio Coding

**Model:** We assume that our audio signal can be modeled as a sum of short sinusoidal oscillations, with a smooth start and a smooth end.

![Example impulse responses of an "MDCT" filter bank](image-url)
We use "modulated" filter banks, meaning the impulse responses are obtained by multiplying (modulating) a baseband prototype (a low pass filter) by a cosine or sine function. The "window" function is the \textit{time-reversed} baseband prototype.

\textbf{Figure:} The baseband prototype or window function of the previous impulse responses
Subband Decomposition: A Filter Bank

Figure: An $N$ - channel filter bank with critical downsampling, perfect reconstruction, and a system delay of $n_0$ samples. Observe: The reconstructed signal is the **weighted sum of the impulse responses**
Filter Banks

- Analysis filter bank generates sequence of short-time spectral representations of a signal
- Number of samples is not increased by the filter bank
- Synthesis filter bank reconstructs signal from this representation
- Perfect reconstruction means $\hat{x}(n - n_d) = x(n)$
- Introduces a system delay $n_d$ from before the analysis to after the synthesis.
Structure of an Audio Coder

Coder/Analysis

Decoder/Synthesis

Signal

Analysis

Psychoacoustics

Q_0

Q_{N-1}

Spec-trum

Synthesis

Decoded signal

Psychoacoustics

Figure: AFB: Analysis filter bank, SFB: Synthesis filter bank
Limits of the Model

Our model of a superposition of impulse responses of a certain length and symmetry assumes some stationarity over the length of the impulse responses, and a similar symmetry of the signals waveform. But: Some audio signals are very short, like attacks of drums or of castanets. Solution: Adapting the model to shorter impulse responses by switching the filter bank to a reduced number of subbands with shorter impulse responses.
Pre-Echos

**Figure:** Pre-echo resulting from too long impulse responses in a Castanets signal
Adapting the Model: Block Switching

“Block length” ~ Impulse response length

Shorter IR, less noise spread in time!

**Figure:** Reducing pre-echo by using block switching
Time-varying filter banks

- can change filters and number of bands during processing of a signal.
- used for non-stationary signals
- adaptable to local signal statistics
- increase compression rate or reconstructed signal quality
Adapting the Model: Switching the Impulse Response

Figure: Sequence of impulse response shapes during switching down and up
Adapting the Model: Non-Symmetric Impulse Responses

- Our "MDCT" impulse responses are symmetric in time,
- but most audio signals are non-symmetric, with fast rise and slow decay, like our Castanet signal.
- Also the ear has a non symmetric sensitivity in time, with high sensitivity before an attack, like of drums, and reduced sensitivity for a short period of time after the attack (order of 100ms)
Temporal Noise Spreading from Symmetric Impulse Responses

**Figure:** A reconstructed impulse after encoding with quantization and decoding. Observe: quantization noise is symmetric around the reconstructed impulse.
Natural Audio Signals Tend to be Non-Symmetric
Non-Symmetric Temporal Masking of the Ear

Figure: The temporal masking threshold of the human ear. Observe the much shorter "pre-masking" duration
Non-Stationary Signals are Difficult to Code with Symmetric Filters

**Figure**: Typical temporal masking threshold and quantization noise of an impulse like signal.
Problem of orthogonal filter banks

- Audible distortions (Pre-echoes) by exceeding the pre-masking level
- Unused masking opportunities.
- ”System delay” $n_d$ from before Analysis to after Synthesis filter bank is filter length -1.
Goal

- Make a filter bank with noise distribution similar to those asymmetric masking curves, with a fast rise and a slow decay.
Approach

- Duration of the noise before the reconstructed Impulse cannot be longer than the system delay ⇒
- Design filter bank where system delay is independent of filter length $L$ to achieve $n_d < L - 1$, low delay.
Design of Filter Banks with Low System Delay

Desired properties of the new filter bank

1. Perfect reconstruction
2. Modulated filter bank with efficient implementation
3. FIR filters for stability
4. System delay independent of filter length
5. Enable time-variability, adaptability to local signal statistics
Approach: Polyphase Representation, Block-wise Processing

\[ x(n) \rightarrow P_a(z) \rightarrow \hat{x}(n) \]

- \( P_a(z) \) Analysis Polyphase Matrix
- \( P_s(z) \) Synthesis Polyphase Matrix
Polyphase representation

Analysis filtering and downsampling becomes a matrix multiplication,

\[ Y(z) = X(z) \cdot P_a(z) \]

\( Y(z) \) is the vector of downsampled \textbf{subband} signals in the z-domain, \( X(z) \) is the vector of the downsampled audio signal at different phases, and \( P_a(z) \) is the ”polyphase matrix” of the analysis filter bank coefficients. On the synthesis side, the upsampling and filtering is

\[ \hat{X}(z) = Y(z) \cdot P_s(z) \]

with \( \hat{X}(z) \) the \textbf{reconstructed} signal.

Hence perfect reconstruction is obtained with

\[ P_s(z) = P_a(z)^{-1} \cdot z^{-d} \]

Observe: We need the \textbf{inverse} matrix, and a \textbf{Delay} \( z^{-d} \) is necessary to make the inverse \textbf{causal}. 

Gerald Schuller
Ilmenau University of Technology and Fraunhofer Institute for Digital Media Technology (IDMT)
Parametric Audio Processing
Obtain FIR filters for Stability and System Delay
Independent of Filter Length

Factorize $F_a(z)$ and $F_s(z)$.

- Construct the matrix $F_a(z)$ and $F_s(z)$ as a cascade of simpler “Filter Matrices” of degree one.
- The Filter Matrices and their inverses must lead to FIR filters.
- The inverse of the Filter Matrices should be already causal to avoid additional system delay.
- 2 independent design parameters, length of the filters, system delay
  $\Rightarrow$ 2 types of Filter Matrices necessary.
Zero-Delay Matrices— increase the filter length by 1 block of $N$ samples, but not the system delay.

$$
L(z) := \begin{bmatrix}
  z^{-1}l_0 & \cdots & & 1 \\
  \vdots & \ddots & \ddots & \ddots \\
  1 & & z^{-1}l_{N/2-1} & 1 \\
  & & 1 & 0 \\
  & & & & 0
\end{bmatrix}
$$
Zero-Delay Matrices

Observe: their inverse is already **causal** and leads to FIR filters,

\[ L^{-1}(z) = \begin{bmatrix}
0 & & & & & 1 \\
\vdots & \ddots & & & & \ddots & \ddots \\
0 & 1 & -z^{-1} l_{N/2 - 1} & & & & \\
1 & \ddots & \ddots & \ddots & & & \\
1 & & & & & \ddots & \ddots \\
& & & & & & -z^{-1} l_0
\end{bmatrix} \]

where \( l_i \) is a real or complex number.
Maximum-Delay Matrices

– they also increase the filter length by 1 block of $N$ samples but the system delay by $2N$.

$$H(z) := egin{bmatrix} h_0 & z^{-1} & \cdots & \cdots & \cdots \\ & h_{N/2-1} & z^{-1} & 0 & \cdots \\ & \cdots & \cdots & \cdots & \cdots \\ & \cdots & \cdots & \cdots & \cdots \\ z^{-1} & \cdots & \cdots & \cdots & 0 \end{bmatrix}$$
Maximum-Delay Matrices - Inverse

Their inverse need a multiplication with $z^{-2}$ to make them causal.

$$H^{-1}(z) \cdot z^{-2} = \begin{bmatrix} 0 & & & & & & z^{-1} \\ & \ddots & & & & \vdots & \vdots \\ & & 0 & z^{-1} & & \ddots & \ddots \\ & & z^{-1} & -h_{N/2-1} & & & \ddots \\ & \vdots & & \ddots & \ddots & & & \ddots \\ & & & & & & -h_0 \end{bmatrix}$$

where $h_i$ is a real or complex number.
The structure can be interpreted as a kind of prediction between different blocks of the signal, to decorrelate them.

Zero-Delay Matrices \( \equiv \) prediction from past samples.
Structure of the Maximum-Delay Filter Matrices

Maximum-Delay Matrices ≡ prediction from future samples, delay for causality necessary.
Optimization

Is an important part of filter bank design,

- Is used to obtain the coefficients of Filter Matrices for the desired magnitude responses
- Only the baseband filters need to be optimized
- A numerical algorithm similar to the method of Conjugate Gradients and the Gauss-Newton Method was developed
- It was found to be robust and fast
- Converges even for big filter banks
Experimental Results

Baseband impulse responses of filter banks with $N = 128$ bands, system delay $n_d = 255$.

- Dashed line: An orthogonal filter bank, identical for analysis and synthesis, filter length $L = 256$
- Solid line: A low delay filter bank, same system delay but longer filters, $L = 512$. 

![Impulse response graph](image-url)
The low delay filter bank has a better magnitude response

The Low Delay filter bank has the **better frequency response** because the prototype can be longer, with same system delay $n_d$.

**Figure**: Frequency magnitude responses for $N = 128$ subbands, lengths $L = 256$ vs. $L = 512$, and same system delay $n_d = 255$. 
The baseband filter commonly used in MPEG coders

Baseband impulse response of an orthogonal filter bank, known as the short sine window.

**Figure:** Sine window for $N = 128$ subbands, length $L = 256$, system delay $n_d = 255$. 
Filter with a lower system delay

Baseband impulse response of a low delay filter bank, similar magnitude response but **lower system delay**. Observe the non-symmetric, minimum phase like appearance.

\[ N = 128, \ L = 512, \ n_d = 191. \]
Design of Time-Varying Filter Banks

This structure can even be used for time-varying filter banks. It only needs a new polyphase formulation for time-varying filter banks

\[ Y(z) = X(z) \cdot P_a(z, m) \]

with time index \( m \) at the lower sampling rate.

\[ \hat{X}(z) = Y(z) \cdot P_s(z, m) \]

**Rule for multiplication** with delay \( z^{-1} \)

\[ P(z, m) \cdot z^{-1} = z^{-1} \cdot P(z, m - 1) \]

- Same structure can be used
- Synthesis Filter Matrices need to be inverse to analysis Filter Matrices at right time

Observe: In an ”overcomplete” representation, long and short filters can also be used in **parallel**, instead of switching, and their coefficients obtained e.g. with ”Matching Pursuit”, for instance for signal analysis.
Temporal Shaped Quantization Noise

Figure: A typical reconstructed impulse with quantization noise using the low delay filter bank.
Filter Bank with High Number of Bands

The MPEG AAC audio codec uses 1024 and 128 bands, switchable.

Figure: Impulse response of the baseband low delay prototype of a filter bank with $N = 1024$ bands length $L = 4096$ taps, and system delay of $n_d = 2047$ samples, identical for analysis and synthesis.
Transition from High to Lower Number of Bands

Figure: Synthesis baseband impulse responses of a low delay filter bank for a transition from 1024 bands to 128 bands.

Gerald Schuller
Ilmenau University of Technology and Fraunhofer Institute for Digital Media Technology (IDMT)
Parametric Audio Processing
Example: MPEG Enhanced Low Delay Audio Coder

Figure: The analysis and synthesis low delay windows of the new MPEG Enhanced Low Delay (ELD) AAC coder.
MPEG ELD-AAC Coder Applications

- Part of the iOS and Android operating systems
- Used in iOS Facetime App
- Achieves an encoding/decoding delay on the order of 10 ms.
- Important for real time audio communications
Example: Musical Instrument Coder

- **Goal:** Synthesize high quality instrument sound with a few easy to estimate parameters
- **Problem to solve:** Finding a suitable model
- **Approach:** Take physical instrument models with parameters that can be estimated in an encoder
Our Example: String Instruments
The String Model

$H(z)$ : Loop filter, modelling the losses
$F(z)$ : Fractional delay filter
$z^{-L}$ : L sample delay
Structure of the Encoder
Structure of the Decoder

- Decoding of the parameters
- Resynthesis with the string model
Coding of the Parameters

- Quantization of the excitation function: 16bit → 8bit
- Quantization of the parameters to 16 bit
- → Bit-rates around 2,4 kbit/s

Subjective listening tests: the decoded signal still sounds somewhat artificial, but without the typical subband based coder artifacts. Observe: These coefficients also have musical meaning, and hence can also be used for classification or to extract muscial information.
Example: Musical Instrument Separation Model

Another example: to separate a musical instrument from a mix, usually time/frequency (spectrogram) models are used.

Figure: The basic structure of a system to separate musical instruments from a mix
Harmonic Model

Most instruments have a **harmonic spectrum, almost**

*Figure*: The harmonics of a trumpet sound. Observe the deviation (lower) from the ideal harmonics
Inharmonicity Coefficient for Strings

String sound harmonics $f_k$ deviate towards higher frequencies from the ideal harmonics $f_0 \cdot k$, because of the stiffness of the string. This is captured on a **model coefficient** $\beta$, which needs to be estimated.

$$f_k = f_0 \cdot k \cdot \sqrt{1 + \beta \cdot k^2}$$
For each instrument tone, we need to estimate the parameters beginning and end times, fundamental frequency, and inharmonicity coefficient $\beta$. Then we can separate them in the time/frequency plane (spectrogram). The spectrogram (with $f$: frequency, $n$: time) of the musical instrument spectrogram is

$$s(f, n)$$

We construct a time/frequency mask

$$M(f, n) = \begin{cases} 
1 & \text{if instrument is in tile } (f, n) \\
0 & \text{else}
\end{cases}$$

The separated musical instrument signal is then obtained from the product spectrogram $M(f, n) \cdot s(f, n)$. 
Conclusions

- We saw a few examples of parametric signal and object models, with emphasis on filter banks for signal representation.
- Parametric object models are used and useful for many applications in audio and image processing.
- For Compression: audio coding, musical instrument coding, 3D model compression
- For signal separation: separate musical instruments from a mix, separate 3-D models from multiple pictures or video signals
- For modification or classification: the extracted parameters can be used to modify or classify the signal in a meaningful way, for instance to classify musical pieces or recognize objects.