Modeling the Spectral Envelope of Musical Instruments

Juan José Burred
burred@nue.tu-berlin.de

IRCAM
Équipe Analyse/Synthèse
Axel Röbel / Xavier Rodet

Technical University of Berlin
Communication Systems Group
Prof. Thomas Sikora

Séminaire Recherche-Technologie
IRCAM, 12th April 2006
Presentation Outline

1. Context: source separation
2. Definition and model requirements
3. Spectral basis decompositions
   - Spectral PCA
   - Previous applications of spectral PCA
   - Training spectral PCA
4. Dealing with variable supports
5. Evaluation framework
6. Experiments and results
7. Modeling of the coefficients
8. Conclusions/future work
### Research context

- **Main research topic:** Underdetermined Source Separation
- **Less mixtures than sources:** strong *a priori knowledge* is needed
  - Knowledge about the mixing process: mixing models
  - Knowledge about the sources
    - General statistic properties: sparsity (past work)
    - Source-dependent modeling (e.g. model of the violin, piano,...)
- **3-month stay at IRCAM** to work on spectral envelope modeling
- **Such a model** will be used in a probabilistic framework as a source of *a priori* knowledge about the signals to be unmixed
- **Other possible applications:** instrument classification, transcription, realistic signal transformations
Spectral envelope: definition

- **Spectral envelope**: a function of frequency that matches the amplitudes of the individual partials of the spectrum.

![Spectral Envelope Graph]

(Figure source: D. Schwarz, “Spectral Envelopes in Sound Analysis and Synthesis”, MSc Thesis, IRCAM, 1998)

- **Motivation**: a sound's spectral envelope is the basic defining factor for its timbre.

- **Dynamic behaviour**: changes over time and can change with f0.
Desirable features of the model for source separation

Ultimate goal: segregation of the overlapping partial peaks in the spectrum

• **Accuracy**
  – The envelope obtained from the model should match the candidate partials as exactly as possible.
  – Time evolution should be reflected in the model.
  – Demanding requirement that is not always necessary in other modeling applications such as classification or retrieval-by-similarity.

• **Generalization**
  – Ability to handle with unknown, real-world mixtures.
  – Need for database training and extraction of prototypes.

• **Compactness**
  – Efficient computation.
  – Together with generality and accuracy, it implies that the model has captured the essential characteristics of the source.
Methods for spectral envelope extraction

• Estimation on whole spectrum
  – Linear Predictive Coding (LPC)
  – Cepstral smoothing
  – Iterative algorithms (True Envelope)

• Estimation based on additive analysis
  – Additive analysis + interpolation between partials
  – Discrete All-Pole (DAP)
  – Discrete cepstrum

• We have chosen to develop a model based on full additive analysis
  – We can use the frequency information for evaluation and parallel modeling
  – It is possible to resynthesize
Sinusoidal Modeling

• A quasi-periodic signal can be modeled by a sum of sinusoids that evolve in amplitude and frequency:

\[ x[n] \approx \hat{x}[n] = \sum_{p=1}^{P[n]} A_p[n] \cos \Theta_p[n] \]

• The instantaneous frequency is the derivative of the total phase:

\[ \Theta_p[n] = \theta_p[n] + 2\pi \sum_{u=0}^{n} f_p[u] \]

• Frame-based processing: STFT → Pitch detection → Partial tracking

\[ \hat{x}_{pl} = (\hat{A}_{pl}, \hat{f}_{pl}, \hat{\theta}_{pl}) \]

• Resynthesis by time interpolation of the parameters
Spectral Basis Decompositions (1)

• General basis expansion signal model:

\[ X = \sum_{i=1}^{N} c_i b_i = BC \]

- \( X \): original data matrix
- \( C \): transformed data matrix (coefficients)
- \( B \): transformation basis. \( B = [b_1, b_2, \ldots, b_N] \) Columns: basis vectors

(e.g.: DFT, STFT, filter banks, wavelets, PCA, ICA, sparse decompositions)

• Application to time-frequency representations:

\( X \) is a t-f representation with \( k = 1, \ldots, K \) spectral bands and \( n = 1, \ldots, N \) time frames, \( N \gg K \)

- Temporal orientation: \( X(n,k) \rightarrow N \times N \) temporal basis
- Spectral orientation: \( X(k,n) \rightarrow K \times K \) spectral basis
Spectral Basis Decompositions (2)

- **Example:** truncated PCA decomposition of a violin t-f representation with first 3 basis

\[
\text{input data} \quad \mathbf{X} = \mathbf{C} \mathbf{B}
\]

- Interpretation as projection into a vector subspace spanned by \( \mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3] \):
Spectral Basis Decompositions (3)

Adaptive transforms applied to spectral basis decomposition:

- **Principal Component Analysis (PCA)**
  - Yields optimally compact representation
  - Main application: dimensionality reduction

- **Independent Component Analysis (ICA)**
  - Yields statistically independent coefficients
  - Main application: Determined Blind Source Separation
  - Independence has proven unnecessary for our representation purposes

  - When applied to a t-f data matrix it is called **Independent Subspace Analysis (ISA)**
  - Main application: Source Separation from single channel

- **Non-negative Matrix Factorization (NMF)**
  - Basis decomposition with non-negativity constraint
  - Has been used to extract features from magnitude spectrograms
  - However, we will work with logarithmic amplitudes → can be negative
 Principal Component Analysis (1)

- **Problem formulation 1:** find the orthogonal directions of maximum variance of a data set
- **Problem formulation 2:** find the reduced-dimension representation of a data set that minimizes the approximation error
- Both problems are equivalent, and their solution is PCA

[Figure source: T. Jehan, “Creating Music by Listening”, PhD Thesis, MIT, 2005]
Principal Component Analysis (2)

- PCA is defined by the linear transformation

\[ \mathbf{Y} = \mathbf{E}^T \mathbf{X} \]

\( \mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K] \) are the unit-length eigenvectors of the sample covariance matrix of the input data:

\[ \Sigma_X = (\mathbf{X} - \mu)(\mathbf{X} - \mu)^T \]

\[ \Sigma_X = \mathbf{E} \mathbf{D} \mathbf{E}^T \]

\( \mathbf{D} \): diagonal matrix of the eigenvalues, sorted in decreasing order:

\[ \mathbf{D} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_K) \quad , \quad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K \]

- input data \( \mathbf{X} \) must be centered: \( \mathbf{X} \leftarrow \mathbf{X} - \mathbf{E} \{ \mathbf{X} \} \)

- the variance of the i-th principal component equals the i-th eigenvalue

- the output data matrix \( \mathbf{Y} \) is uncorrelated

- PCA can be efficiently implemented with Singular Value Decomposition (SVD)
Principal Component Analysis (3)

- Dimensionality reduction with PCA:
  - keep the first $R < K$ eigenvectors corresponding to the $R$ largest eigenvalues
    \[
    Y_r = E_r^T X
    \]
    $Y_r$: $R \times N$ reduced dimension representation
    $E_r$: $K \times N$ reduced PCA basis
  - approximate reconstruction:
    \[
    \hat{X} = E_r Y_r = E_r E_r^T X
    \]
  - reconstruction error (Mean-Square Error)
    \[
    MSE = E\{\|X - \hat{X}\|^2\}
    \]
    - the MSE is equal to the sum of the ignored eigenvalues
    \[
    MSE = \sum_{i=R+1}^{K} \lambda_i
    \]
An example of spectral PCA

- PCA applied to the partial amplitudes of a single horn note
Previous applications of spectral PCA (1)

- Data reduction of additive analysis/synthesis data [Sandell & Martens, 1995]
  - Perceptual experiments
  - Single notes, no training
  - 40-70% data reduction to obtain nearly identical tones

- Additive analysis/synthesis using Multidimensional Scaling (MDS) [Hourdin, Charbonneau, Moussa, 1997]
  - MDS similar concept to PCA
  - Main goal: representation of sound trajectories in timbre space
  - No training
  - 75% of information for musically acceptable sounds
  - 90% of information for sounds indistinguishable from the original
Previous applications of spectral PCA (2)

- **Sonological models** for timbre characterization [De Poli & Prandoni, 1997]
  - PCA input data are a fixed number of MFCC cepstral coefficients
  - Rough approximation of the envelope, no training

```
(Figure source: G. De Poli, P. Prandoni, “Sonological models for timbre characterization”, J. New Music Research, 1997)
```

- Feature extraction in the MPEG-7 standard [Casey, 2001]
  - Another context: general sound description. Not based on spectral envelope.
Training spectral PCA

- Training database
  - Concatenation of sound samples
  - Pre-processing
    - Centering
      - PCA
        - Model space
          - Coefficient modeling
            - Model database
              - statistical model / curve prototype

- mean
- basis
Dealing with variable supports (1)

- We wish to concatenate the partial amplitudes of several notes in order to train a common PCA basis.
- It is straightforward to extract a fixed number of partials for each training sample and arrange them in the data matrix $X(p,l)$, where $p$ is the partial and $l$ the frame index.

(Partial Indexing, PI)

$$x[n] \approx \hat{x}[n] = \sum_{p=1}^{P[n]} A_p[n] \cos \Theta_p[n]$$

$$P[n] = P$$

$$x_{pl} = \hat{A}_{pl}$$
Dealing with variable supports (2)

- However, when using notes of different pitches to generalize the model we are in effect misaligning some frequency information.

Ex.: Training 1 octave (C4-B4) of an alto saxophone

$X(p,l)$
Dealing with variable supports (3)

• To correct the misalignment of frequency-invariant features (fixed formants, resonances): set maximum frequency → extract a different number of partials for each note → interpolate in frequency to get data matrix (Envelope Interpolation, EI)

• We define a regular frequency grid (grid index: $g$)

• We compare two interpolation methods:
  
  • Linear interpolation:

  $p_0 < g < p_1 \quad A_{gl} = A_{p_0l} + \frac{A_{p_1l} - A_{p_0l}}{f_{p_1l} - f_{p_0l}} (f_g - f_{p_0l})$

  • Cubic polynomial interpolation:

    • Find interpolation polynomial

    $p(f) = a_0 + a_1 f + a_2 f^2 + a_3 f^3$

    so that $p(f_{p_i l}) = A_{p_i l}$
Dealing with variable supports (4)

Ex.: Training 1 octave (C4-B4) of an alto saxophone, extracting all partials up to the 20th partial of the highest note, linearly interpolating with a regular frequency grid of 40 points.
Partial indexing vs. Envelope Interpolation

- Taking the partial index as spectral index in the data matrix misaligns the frequency-invariant features (formants, resonances) of the spectral envelope.
- Frequency interpolation avoids this but introduces interpolation errors.
- On the other hand, partial indexing aligns f0-correlated resonances.
- In principle, frequency alignment is desirable because:
  - Prototype spectral shapes will be learned more effectively.
  - The data matrix will be more correlated and thus PCA will be able to achieve a better compression.
- The question arises:
  - Which of these strategies is more appropriate for the PCA model?
- In other words:
  - What kind of features (f0-correlated or invariant) are more important for our model?
Cross-validation framework

Training database

Preprocessing

PCA/ dim.red

parameters

basis $E_r^T$

Model space

EXP 1: compactness

Test database

Preprocessing

Model space

EXP 2: accuracy

EXP 3: generalization

$E_r$

Reconstruction

Reinterpolation
Results 1: compactness

- Explained, accumulated variance (eigenvalues):

\[
EV(d) = 100 \frac{\sum_{i}^{d} \lambda_i}{\sum_{i}^{D} \lambda_i}
\]

Exp: 4\textsuperscript{th} octave, 2 instr. from the RWC database

Graphs showing the explained variance for piano, bassoon, and violin.
Results 2: Accuracy

- Relative Spectral Error (RSE) of the reconstructed partials, reinterpolated at the original frequencies

\[
RSE = \frac{1}{L} \sum_{l=1}^{L} \sqrt{\frac{\sum_{p=1}^{P_l} (A_{pl} - \tilde{A}_{pl})^2}{\sum_{p=1}^{P_l} A_{pl}^2}}
\]

Exp: 4\textsuperscript{th} octave, Training: 2 instr., Test: 1 instr. from RWC

- Plots showing RSE for bassoon, violin, and piano, comparing PI, linear EI, and cubic EI methods.
Experiment 3: generality (1)

- Problem: measure distance between data cloud of training coefficients and data cloud of test coefficients without assuming any probability distribution.

- The data clouds do not necessarily form a gaussian cluster.

- In such a case, we cannot trust a distribution measure based on normal parameters (divergence, Bhattacharyya, Cross Likelihood Ratio).
Experiment 3: generality (2)

- Measures not assuming any distribution (i.e., solely based on point topology) will be more reliable in the general case.

- **Ex.:** Kullback-Leibler Divergence:

\[
KL(N_0, N_1) = \frac{1}{2} \left( \log \left( \frac{\det \Sigma_1}{\det \Sigma_0} \right) + \text{tr} \left( \Sigma_1^{-1} \Sigma_0 \right) + (\mu_1 - \mu_0)^T \Sigma_1^{-1} (\mu_1 - \mu_0) - N \right)
\]

- Compared to averaged minimum Mahalanobis distance between points:

\[
D(\omega_1, \omega_2) = \frac{1}{n_1} \sum_{i=1}^{n_1} \min_j \{d_M(x_i, x_j)\} + \frac{1}{n_2} \sum_{j=1}^{n_2} \min_i \{d_M(x_i, x_j)\}
\]

where

\[
d_M(x_0, x_1) = \sqrt{(x_0 - x_1)^T \Sigma^{-1} (x_0 - x_1)}
\]
Results 3: generality

- Averaged minimum Mahalanobis distance between training and test data clouds

Exp: 4th octave, Training: 2 instr., Test: 1 instr. from RWC
Modeling the coefficients (1)

• Further generalization is possible by modeling the transformed coefficients

• Common approaches from Music Information Retrieval:
  – GMM (Gaussian Mixture Models)
  – HMM (Hidden Markov Models)

• To fully characterize the dynamic behavior of the envelopes, we choose to model the coefficients as a prototype trajectory.
Modeling the coefficients (2)

- First experiments: simple time interpolation and averaging in low dim space

Exp: piano, 4th octave, Training: 2 instr. from RWC
Modeling the coefficients (3)

- **Example:** training of several instruments on the same space (e.g. for timbre characterization, blind source separation)
Modeling the coefficients (4)

- Further refinement: application of **Principal Curves**
  - Nonlinear extension to PCA
  - Has been used to model gestures captured by sensors

Conclusions / Future work

• When training the PCA model with notes of different pitch, frequency interpolation improves accuracy of the model.

• The interpolation error is compensated by the gain in correlation between envelope time frames in training data.

• Appropriate framework for dynamic timbre modeling using prototype trajectories.

• Future work
  – Integration in a source separation framework
  – Refinement of trajectory models
  – Modeling of frequency information