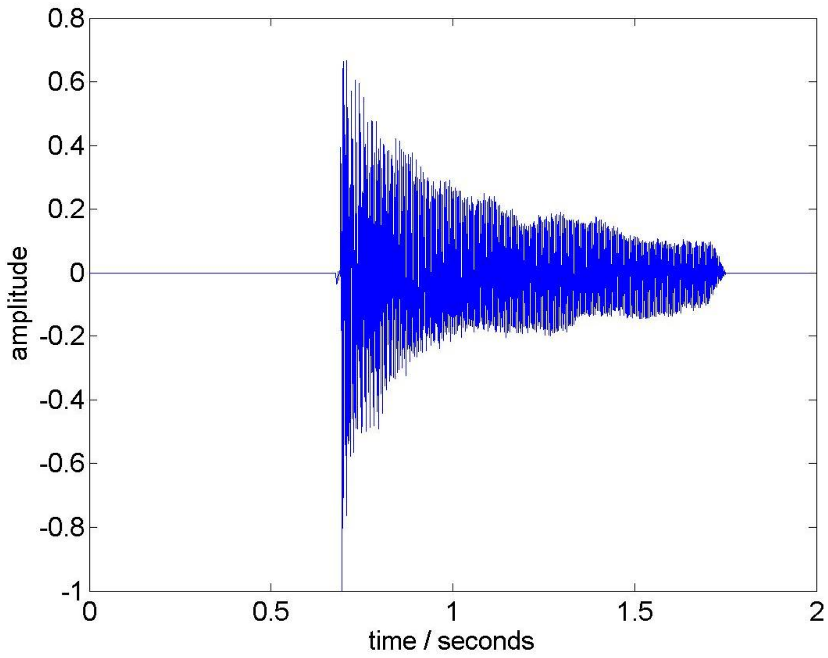
Non-Negative Matrix Factorization And Its Application to Audio

Tuomas Virtanen
Tampere University of Technology
tuomas.virtanen@tut.fi

Contents

Introduction to audio signals

- Audio signal: representation of sound
- Can exist in different forms
  - Acoustic (that's how we hear and often produce it)
  - Electrical voltage (ouput of a microphone, input of a loudspeaker)
  - Digital (mp3 files, compact disc, mobile phone)

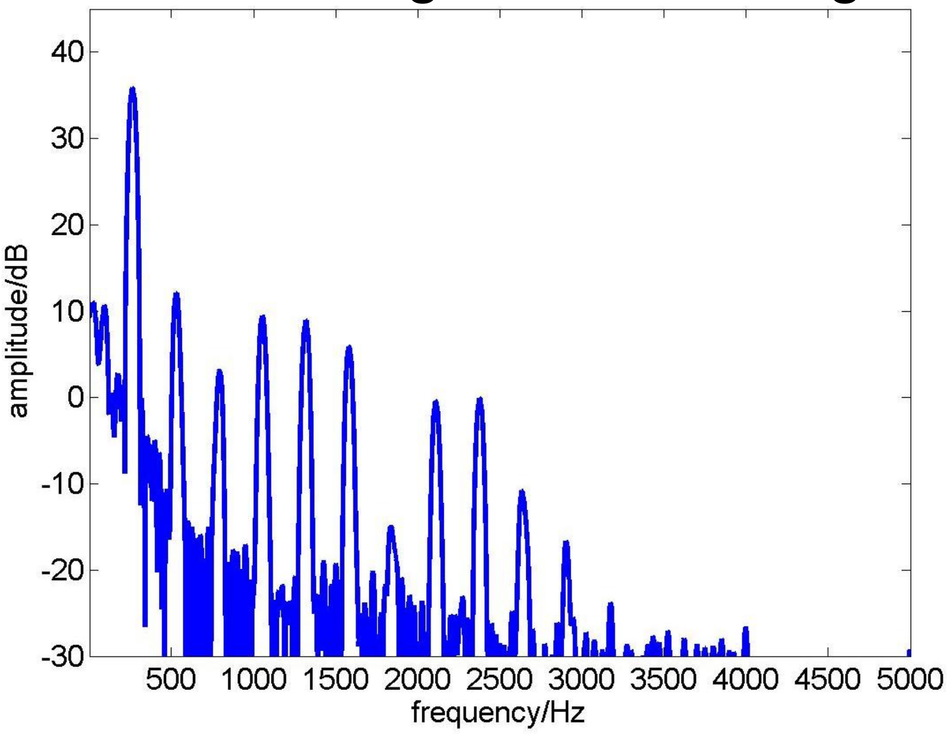


Representations of audio signals

- The amplitude as a function of time is a natural representation of ausio signals
  - Describes the variation of the sound pressure level around the DC
  - Easy to record using a microphone and to reproduce by a loudspeaker
- Digital signals: sampling frequency 44.1 kHz commonly used
  - Allows representing frequencies 0 – 22.05 kHz
  - Humans can hear frequencies 20 Hz-20 kHz
  - Lower / higher sampling frequencies also used
  - Most of the information in low frequencies

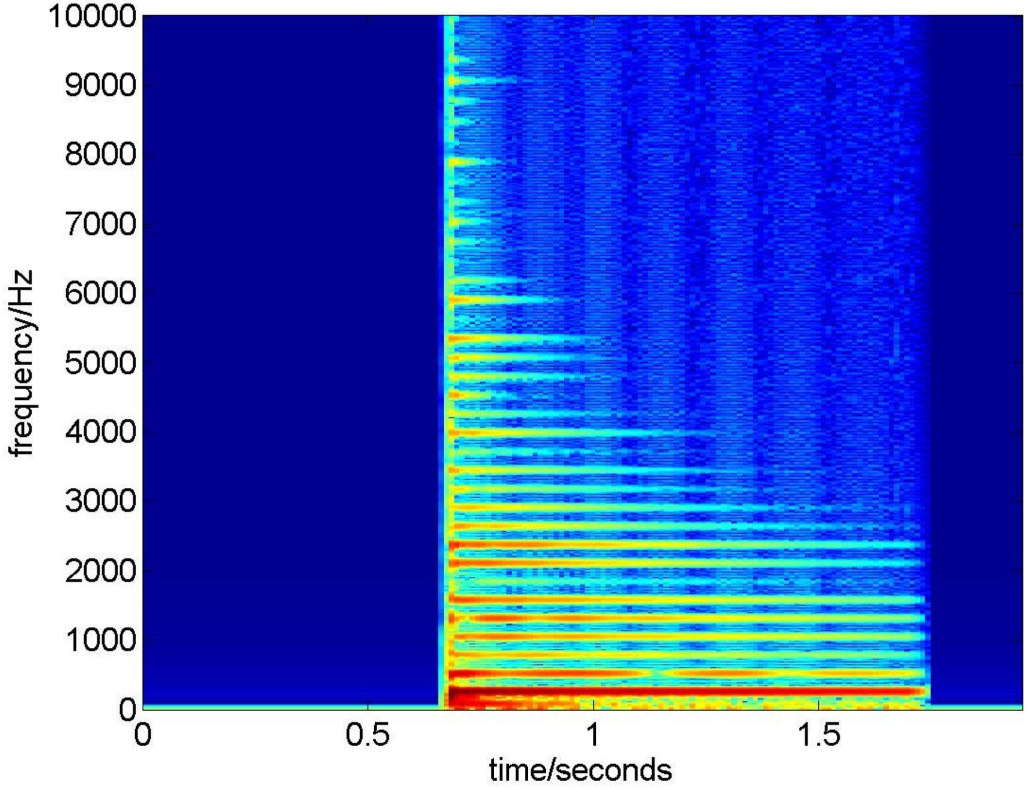
Spectrum of a sound

- Obtained e.g. by calculating the DFT of the signal
- Perceptual properties of a sound are more clearly visible in the spectrum
- Amplitude in dB – closer to the loudness perception
- Phases less meaningful – often magnitudes only are used



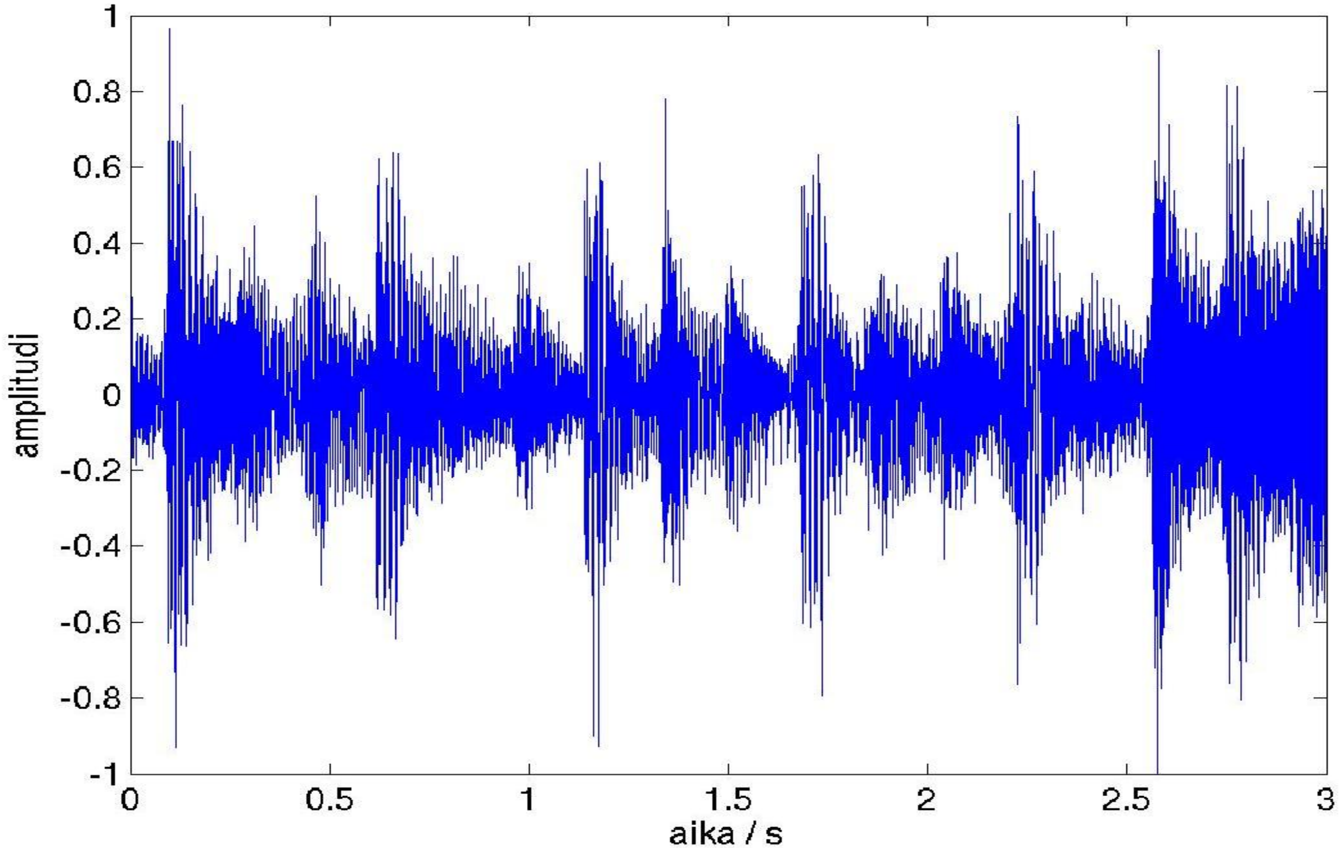
Spectrogram representation

- Represents the intensity of a sound as a function of time and frequency
- Obtained by calculating the spectrum in short frames (10-50 ms typically in the case of audio)



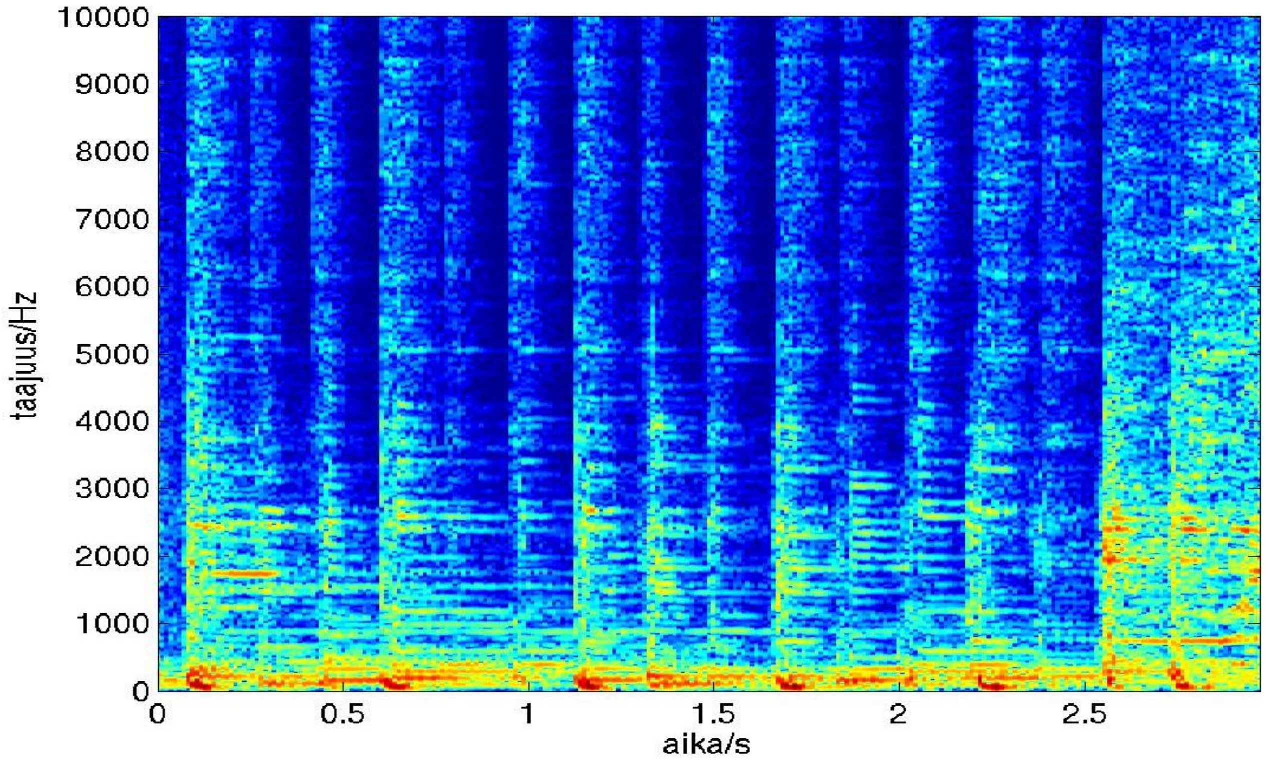
Linear superposition

- When multiple sound sources are present, the signals add linearly



Spectrogram of polyphonic music

- Mid-level representation suitable for audio analysis (Ellis & Rosenthal 1998)
- The rhythmic structure is still visible



Source separation

- In practical situations other sounds interfere the target sound
- Automatic recognition / processing of sounds within mixtures extremely difficult
- Applications:
  - Robust speech recognition
  - Speech enchacement
  - Music content analysis (transcription, instrument identification, singer identification, lyrics transcription)
  - Audio manipulation
  - Object-based coding
- Very important in many other fields

How to separate

- Prior information about sources
- General assumptions: statistical independence, etc.
- Multiple microphones: direction of arrival
- How does the human auditory system separate sources?

Blind source separation

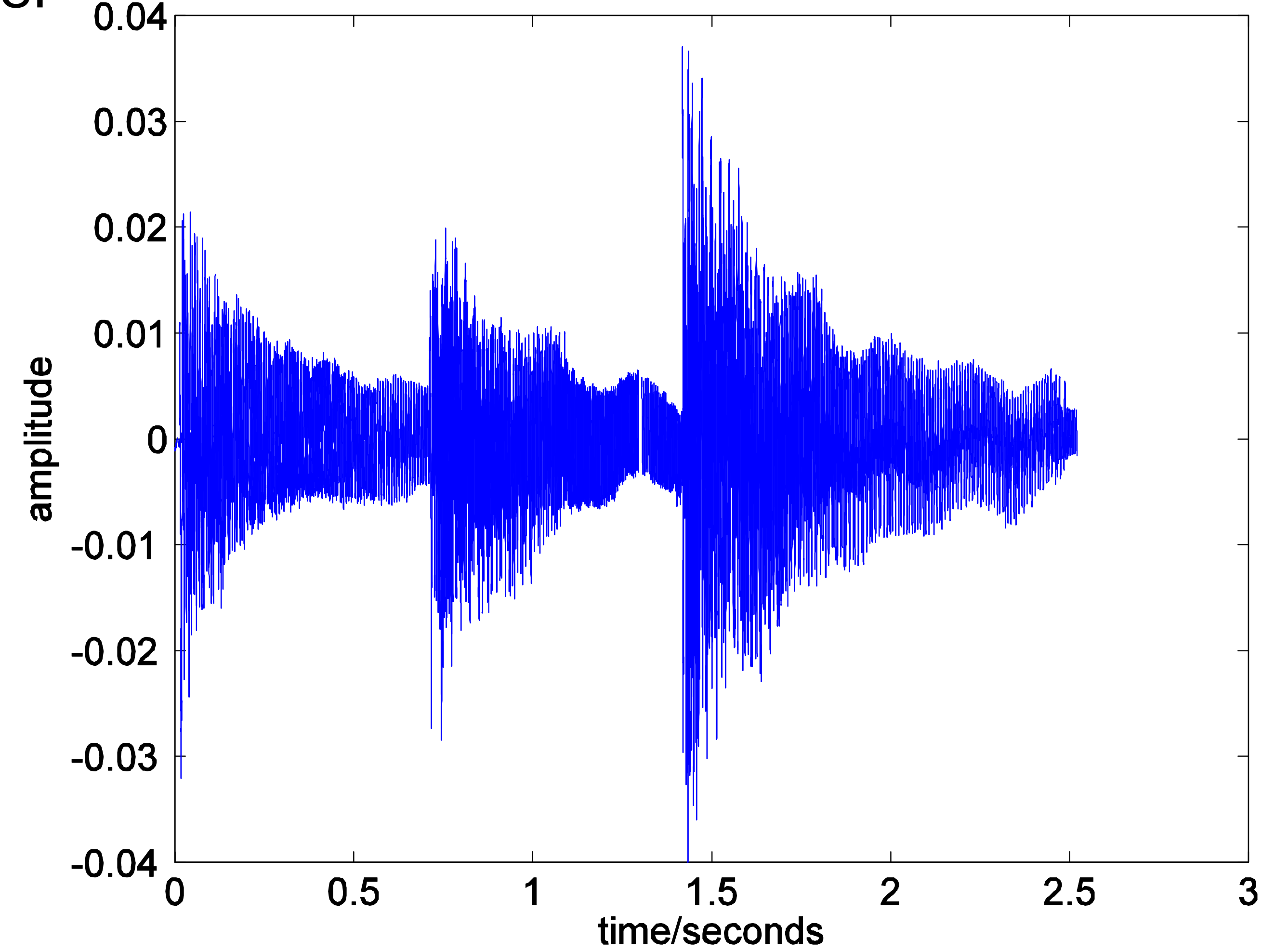
- No prior information about sources
- Only generic assumptions that are valid for all the possible sources
  - E.g. statistical independence
- Involves unsupervised learning
- In many practical situations we have less sensors than sources:
  - How to to estimate multiple signals from a smaller amount of observations?

Sparseness in broad sense

- Assumption: a source signal can be described using a small number of parameters in some domain
- One possible approach: latent variable decompositions

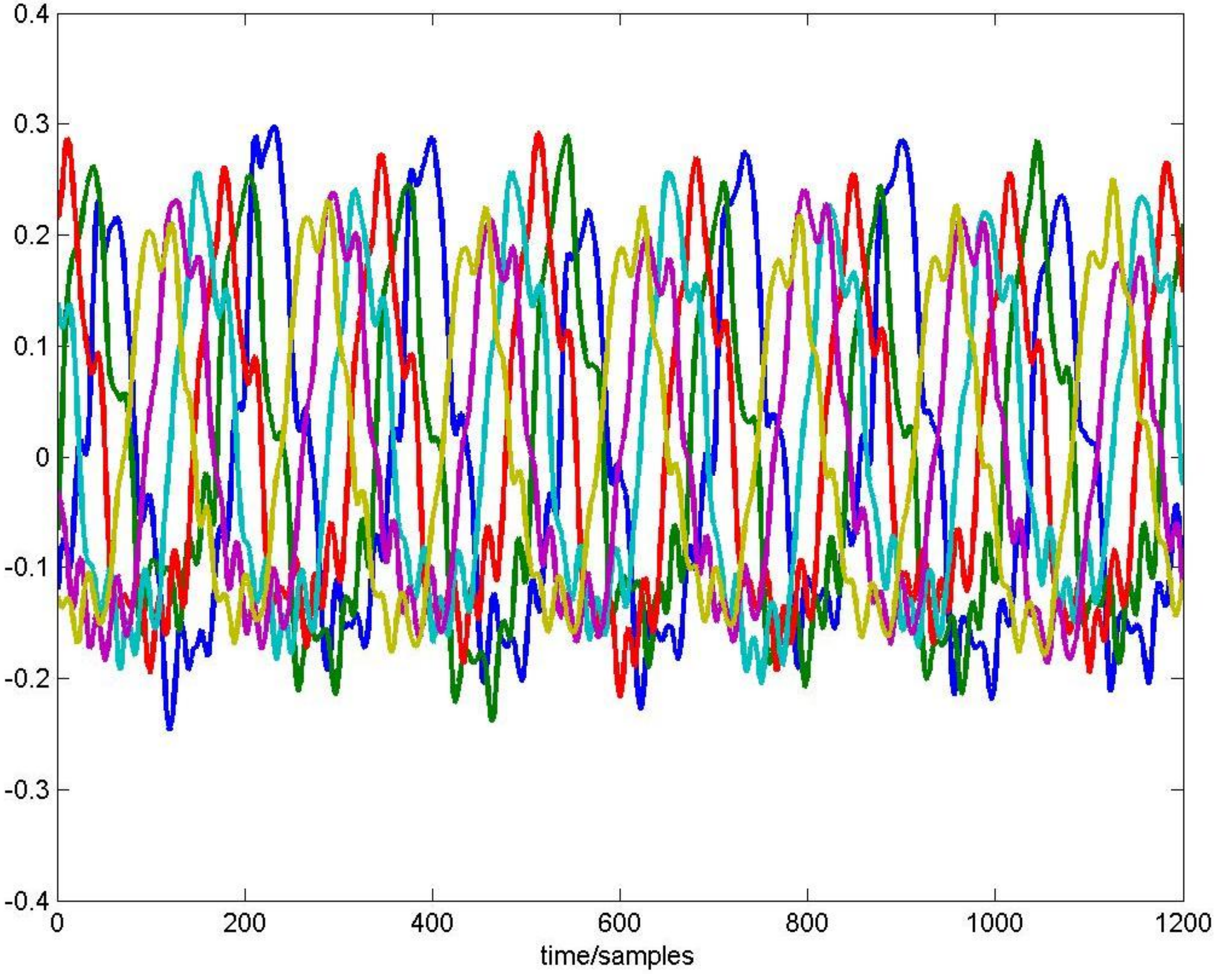
Example signal

- Notes C4 and G4 played by guitar, first separately and then together



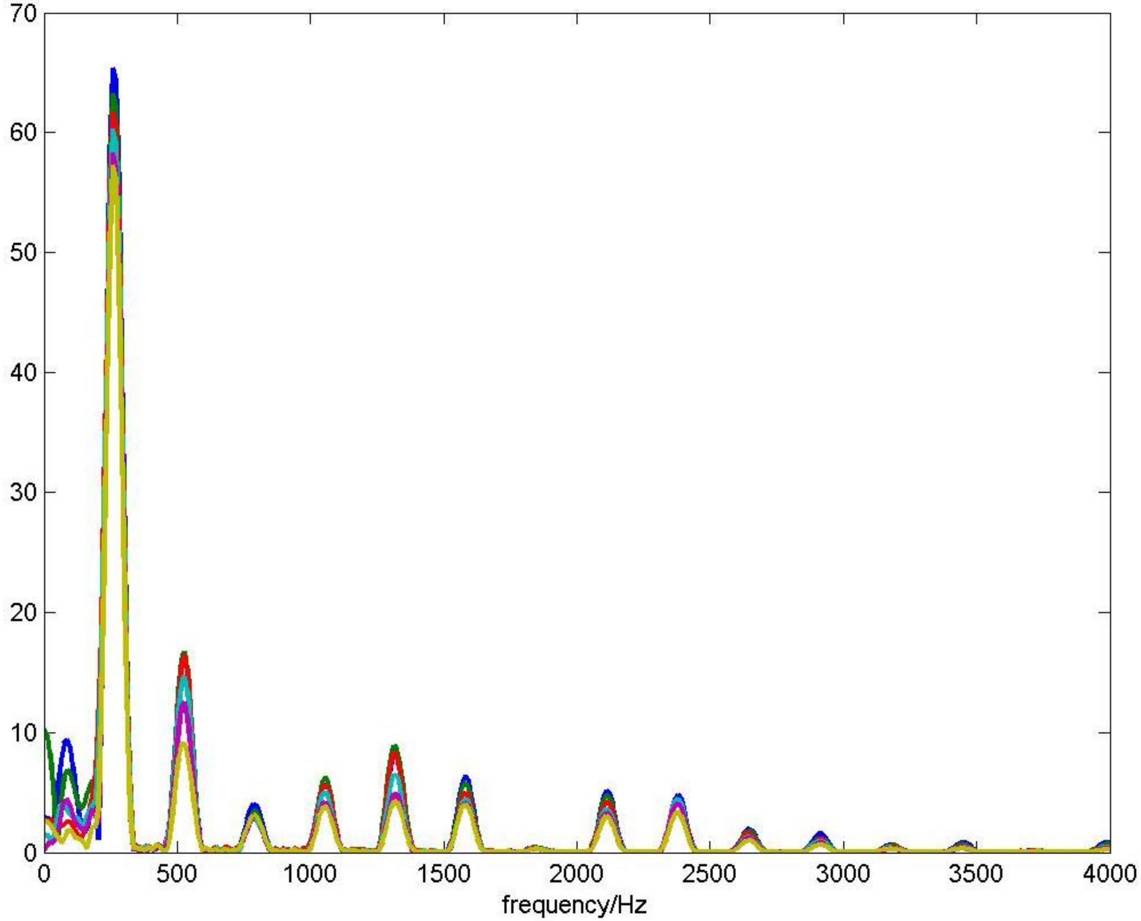
Sparseness of the time-domain signal

- Five frames of the first note:

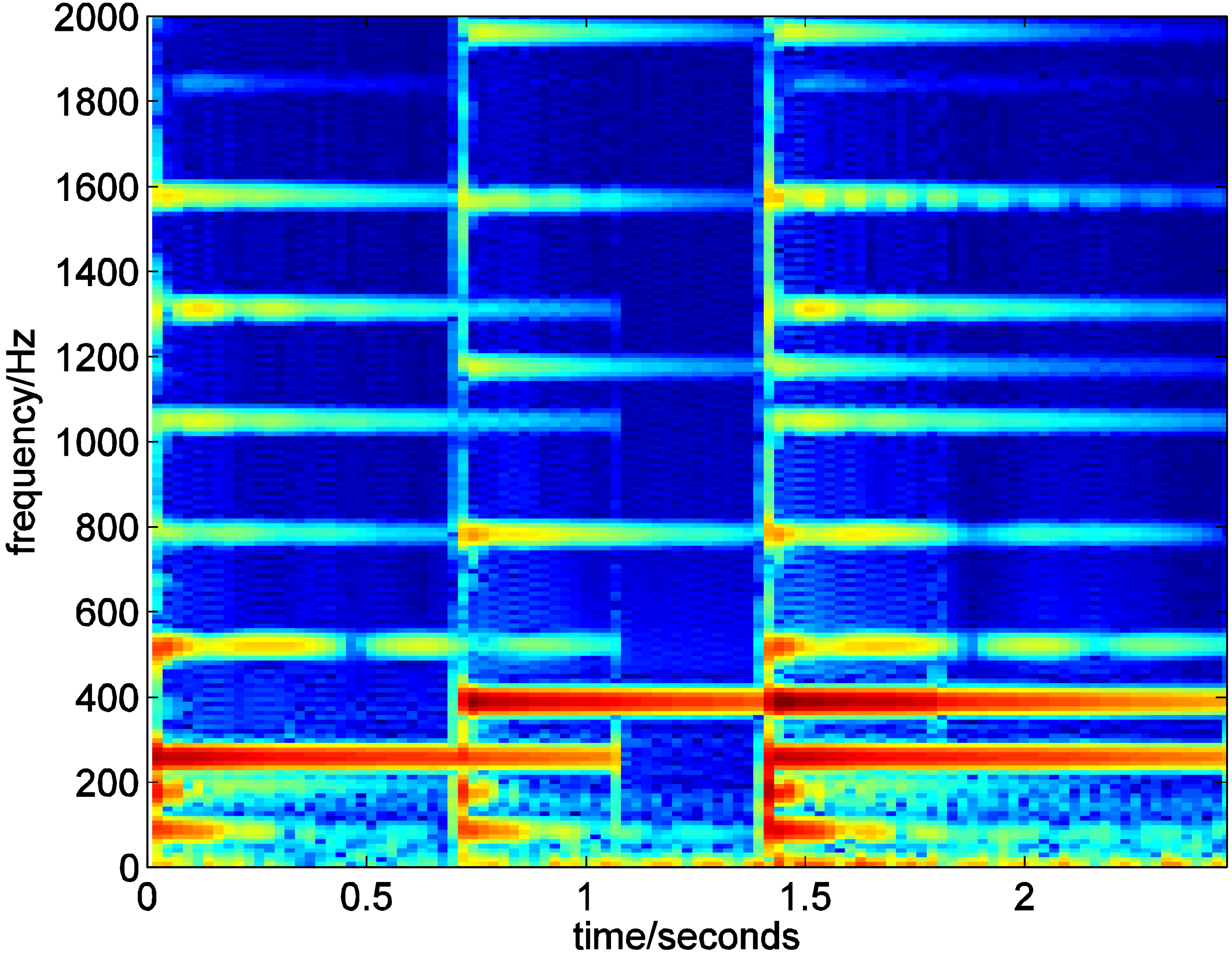


Sparseness of magnitude spectrum

- Five magnitude spectra of the first note: phase-invariant representation leads to much more compact models



Mixture spectrogram



Linear model for the mixture

- Spectrum vector $\mathbf{x}_t$ is decomposed into weighted sum of frequency basis vectors $\mathbf{a}_1$ and $\mathbf{a}_2$

$$\mathbf{x}_t = \mathbf{a}_1 s_{1t} + \mathbf{a}_2 s_{2t}$$

- $\mathbf{a}_1$ and $\mathbf{a}_2$ represent the spectra of note 1 and 2, respectively
- s_{1t} and s_{2t} represent the gain of the notes over time
- Model in vector-matrix form:

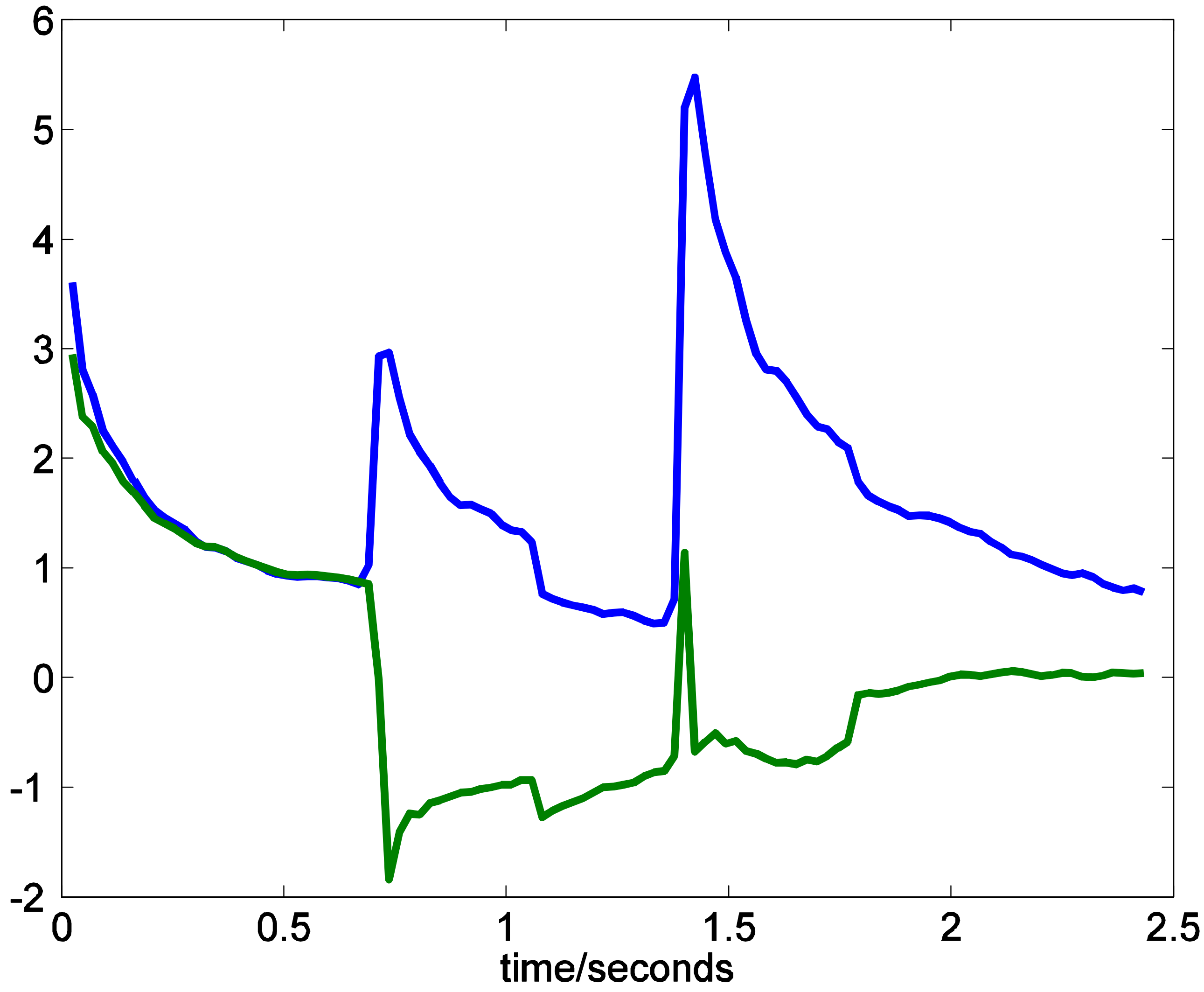
$$\begin{bmatrix} x_{1t} \\ x_{2t} \\ \vdots \\ x_{Ft} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ \vdots & \vdots \\ a_{F1} & a_{F2} \end{bmatrix} \bullet \begin{bmatrix} s_{1t} \\ s_{2t} \end{bmatrix} \qquad \mathbf{x}_t = \mathbf{A}\mathbf{s}_t$$

ICA on spectrogram

- The model matches the ICA model: each frequency is an sensor, mixture weights are sources
- Let us try to use ICA to separate the notes
- ICA on spectrogram: Independent subspace analysis ISA, (Casey & Westner 2000)

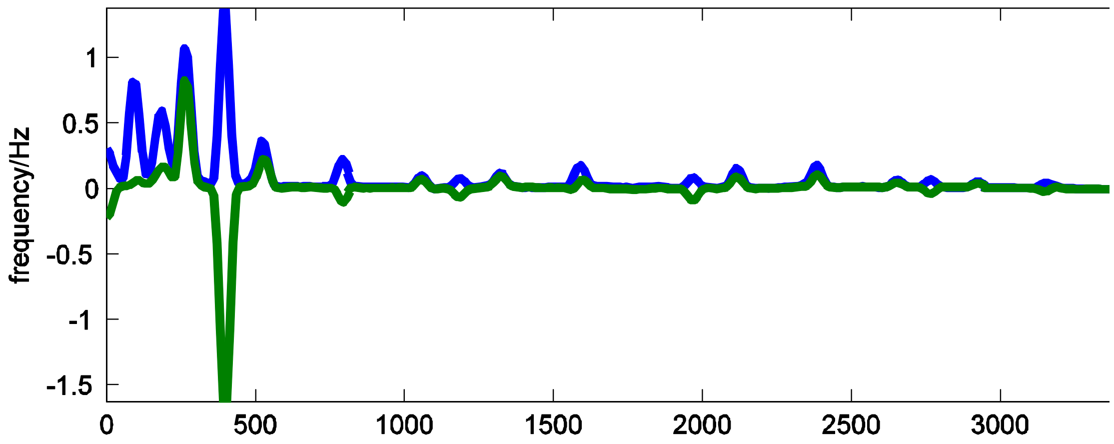
Results with ICA

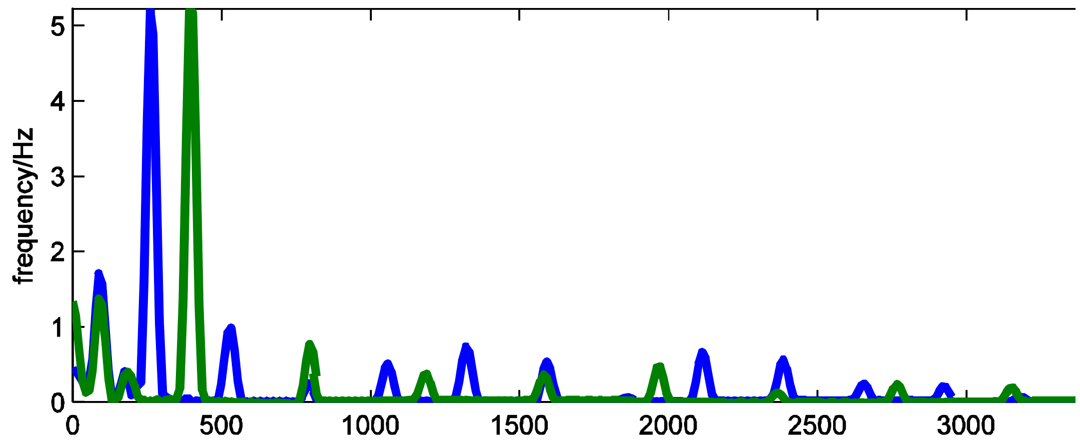
- Weights over time
- Negative weights(!)
- Both weights seem to represent the first note



Spectral basis vectors obtained with ICA

- ICA estimate (upper panel) vs. original (lower panel)
- Both components represent note a combination
- Negative values





What goes wrong?

- Negative weights: subtraction of spectral basis vectors
- Negative values in spectral basis vectors
- Subtraction of magnitude of power spectra physically unrealistic
- Are the notes statistically independent?
- Are the modeling assumptions correct?
- Is the independence as defined in ICA a good assumption in this case?

Non-negativity restrictions

- Non-negativity restrictions difficult to place into ICA
- It has been shown that with non-negativity restrictions, PCA leads to independent components (Plumbley 2002, Wilson & Raj 2010)

Non-negativity restrictions alone

- What if we seek for a representation

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t$$

while restricting the basis vectors and weights to non-negative values?

Model for multiple frames

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t, \quad t = 1, \ldots T$$

written for all the frames in matrix form:

$$\begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_T \end{bmatrix} = \mathbf{A} \bullet \begin{bmatrix} \mathbf{s}_1 & \mathbf{s}_2 & \cdots & \mathbf{s}_T \end{bmatrix}$$

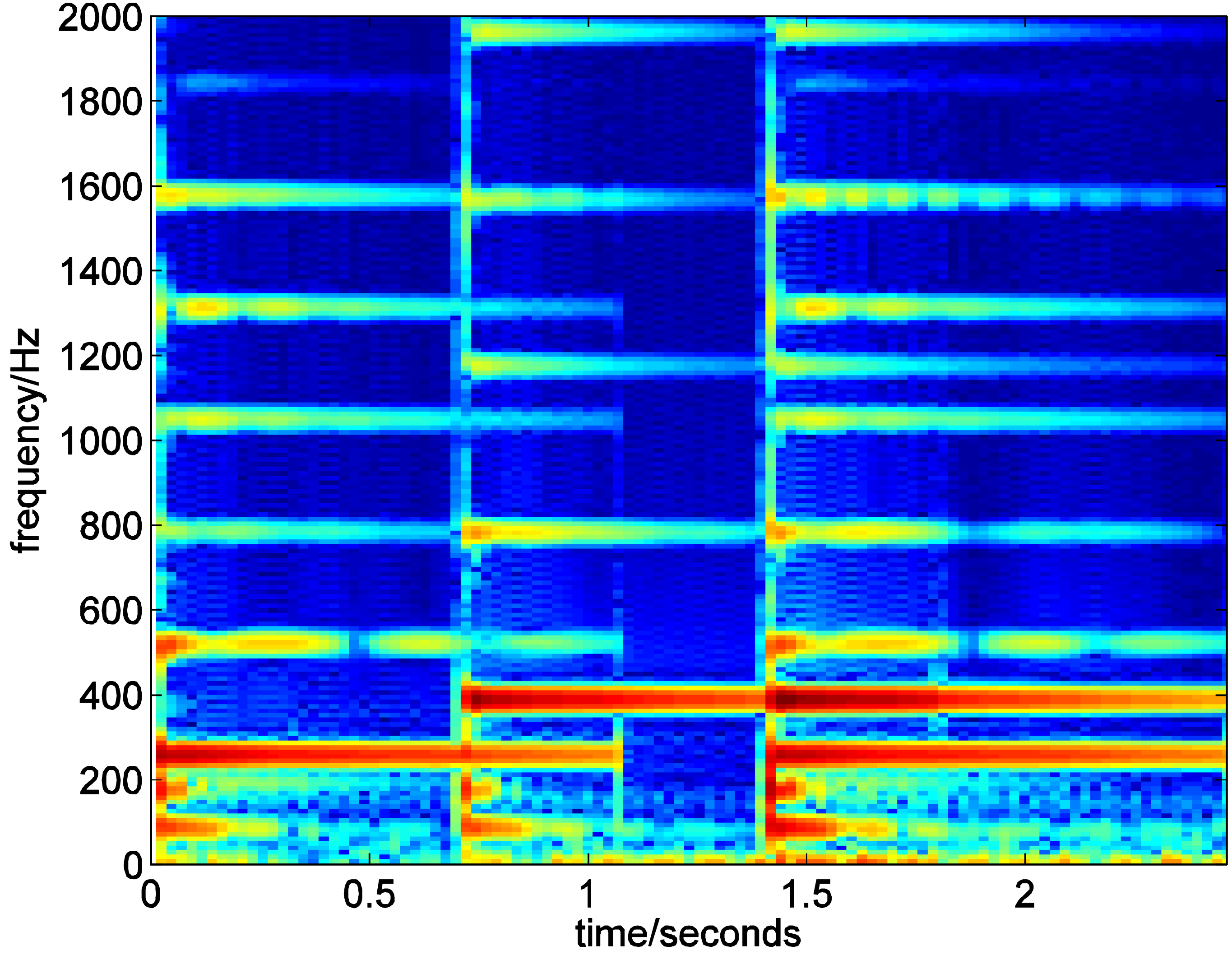
and using matrices only:

$$\mathbf{X} = \mathbf{A}\mathbf{S}$$

Non-negative matrix factorization

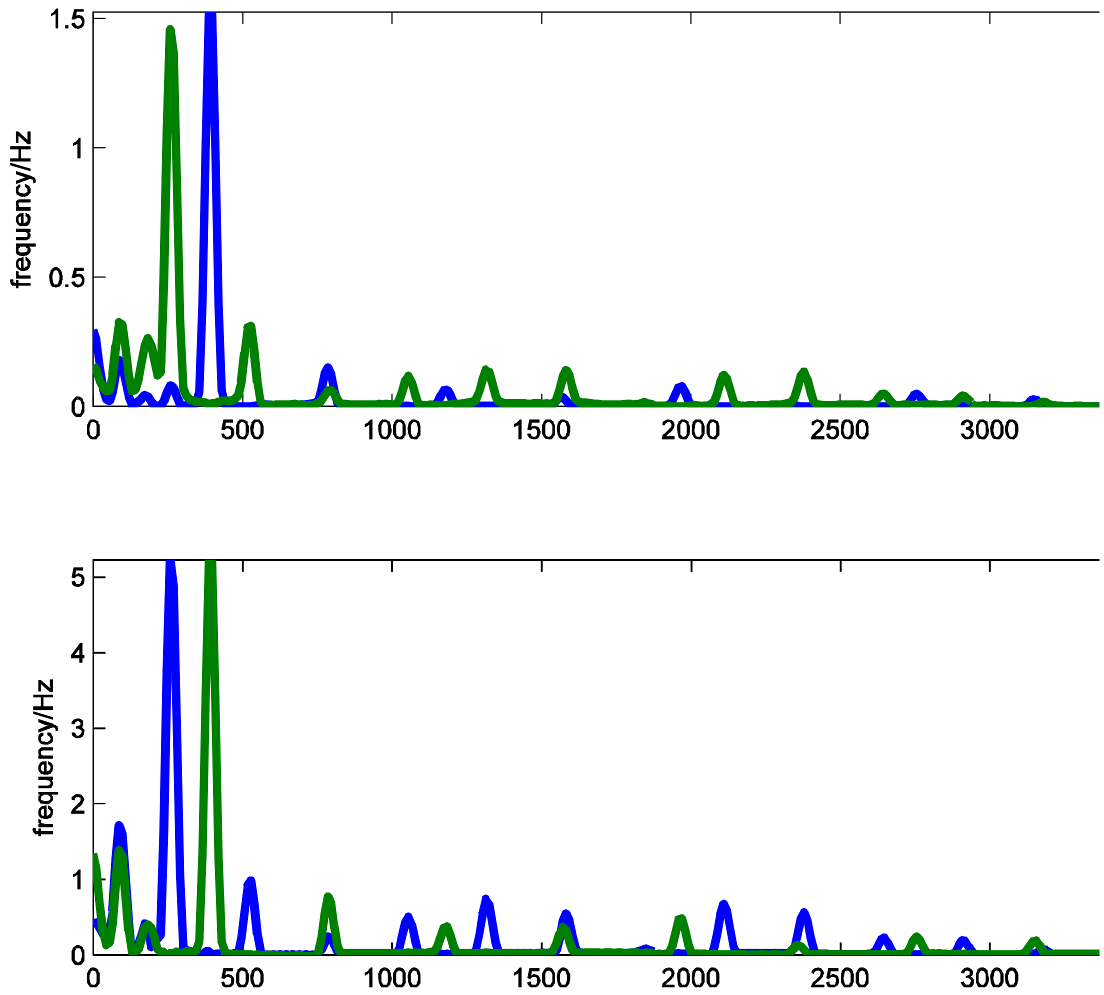
- NMF: minimize the error of the approximation **X** = **AS**, while restricting **A** and **S** to non-negative values (Lee & Seung, 1999 & 2001)

Guitar example



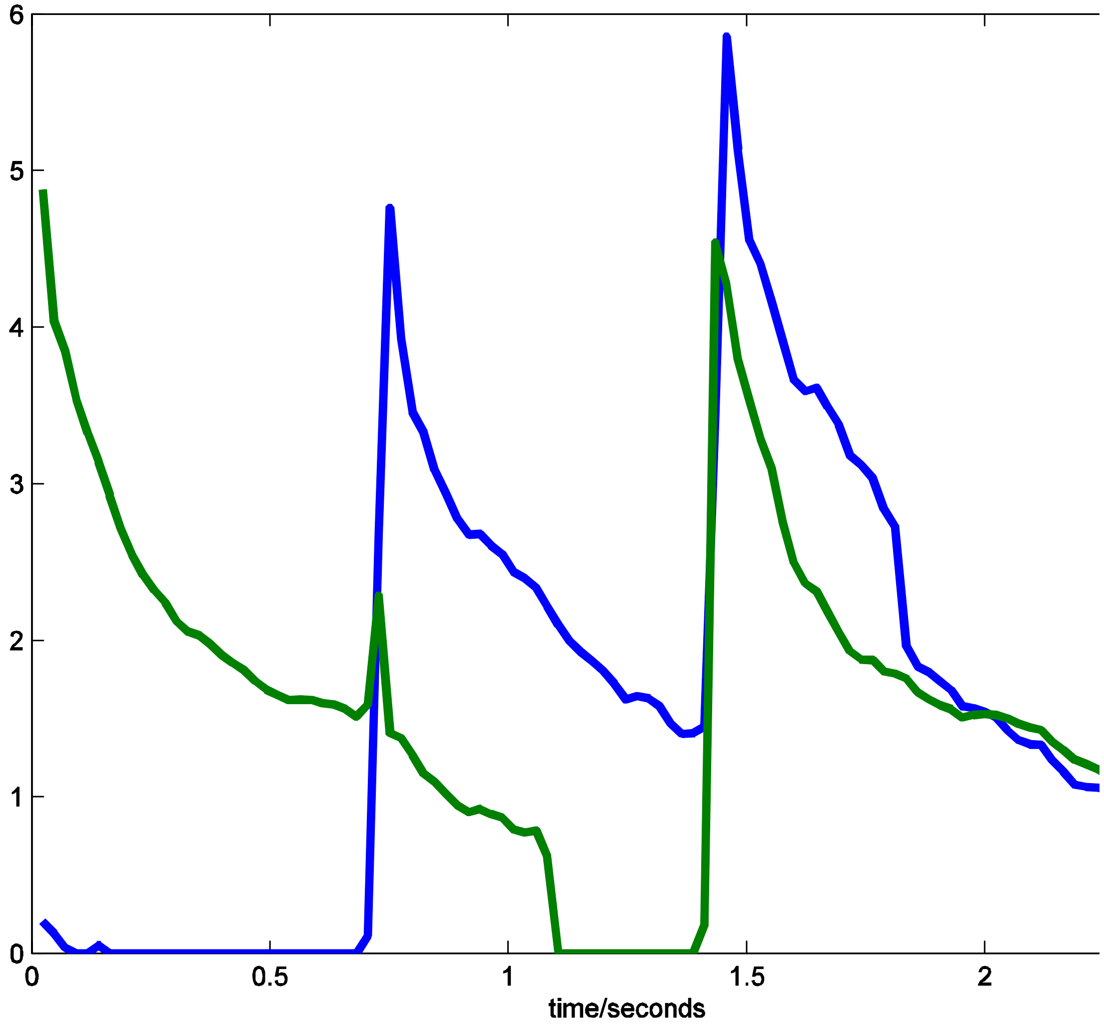
Spectral basis vectors obtained with NMF

- NMF estimate (upper panel) vs. original (lower panel)
- Bases correspond to individual notes
- Permutation ambiquity



Weight obtained with NMF

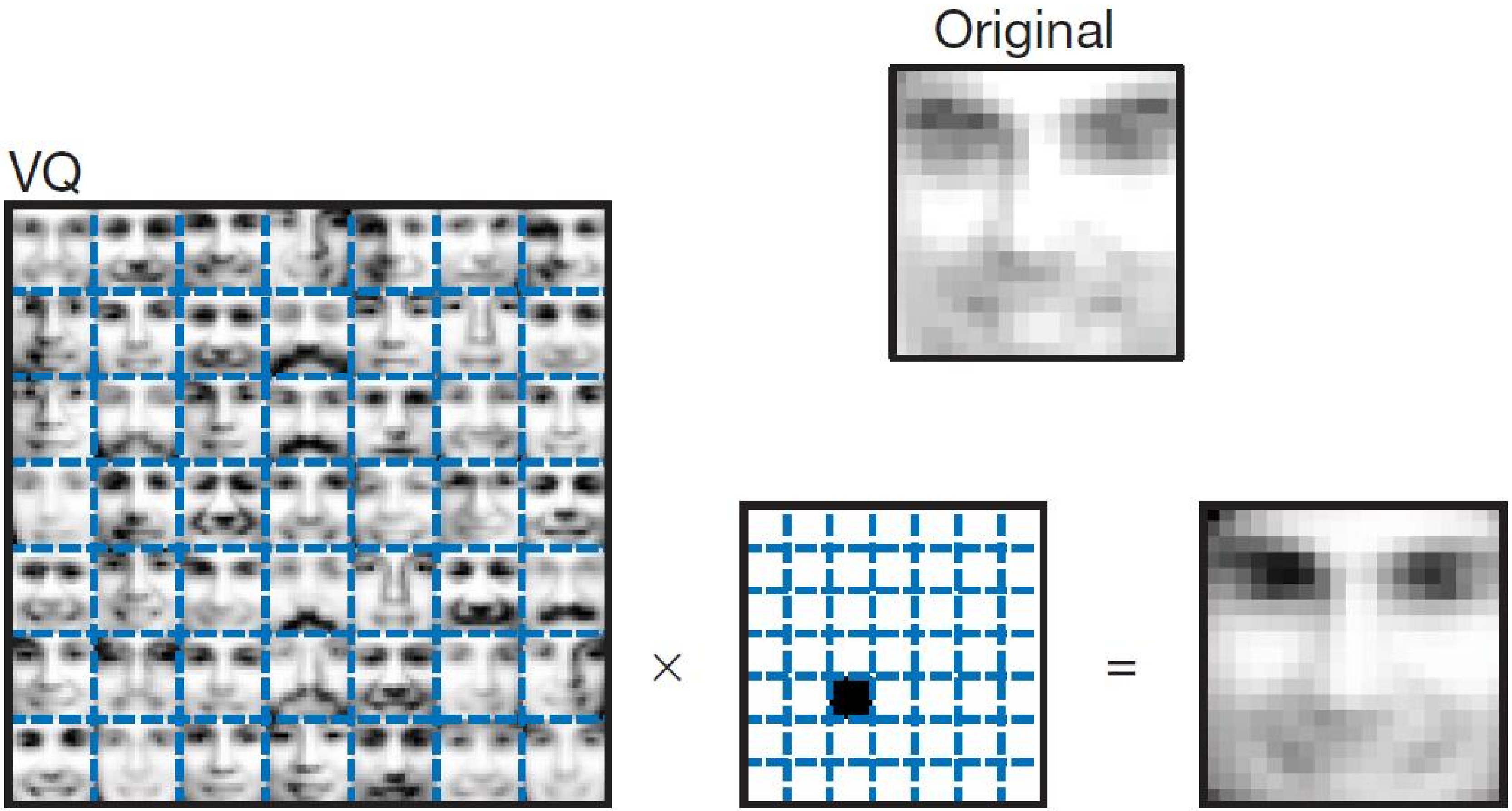
- The green basis represents partly the onset of the second note
- Good separation of notes



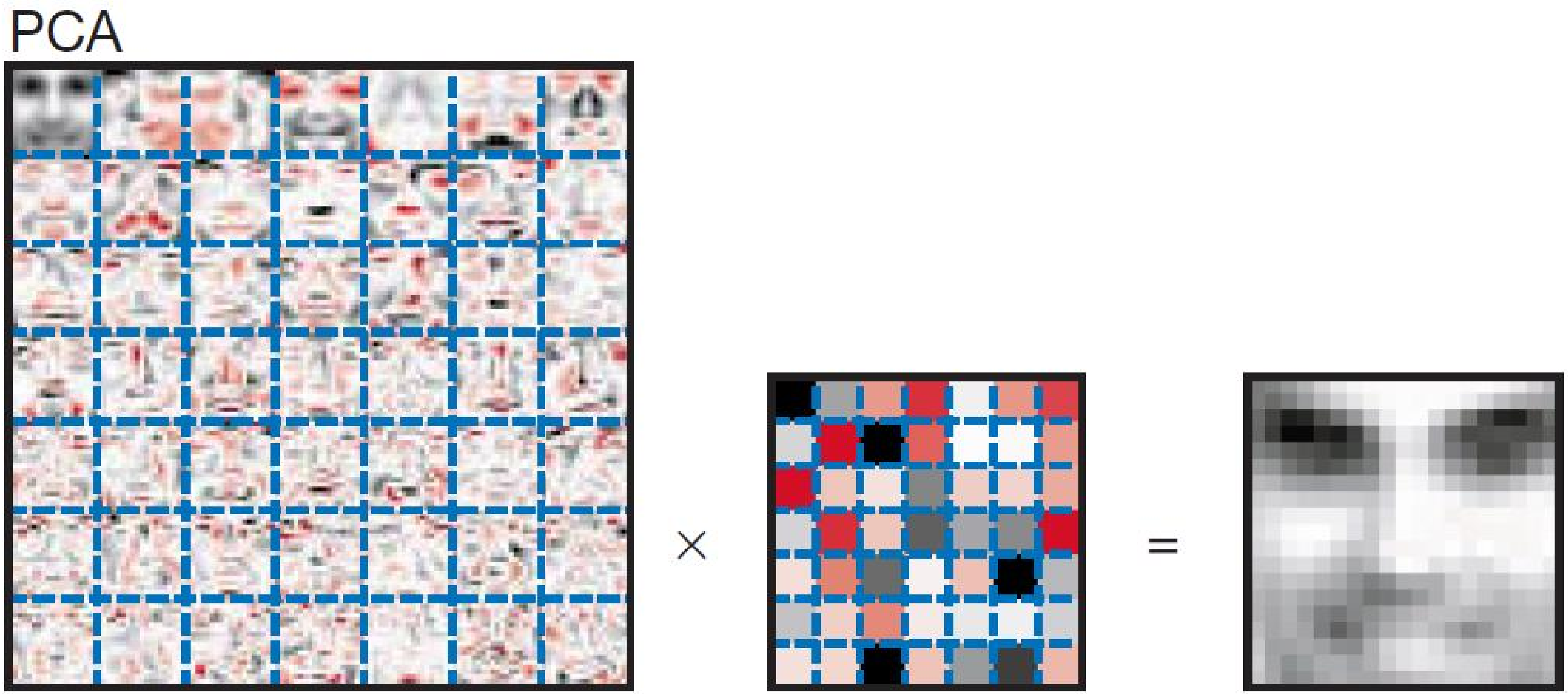
Why does NMF work?

- By representing signals as a sum purely additive, non-negative sources, we get a parts-based representation (Lee & Seung, 1999)

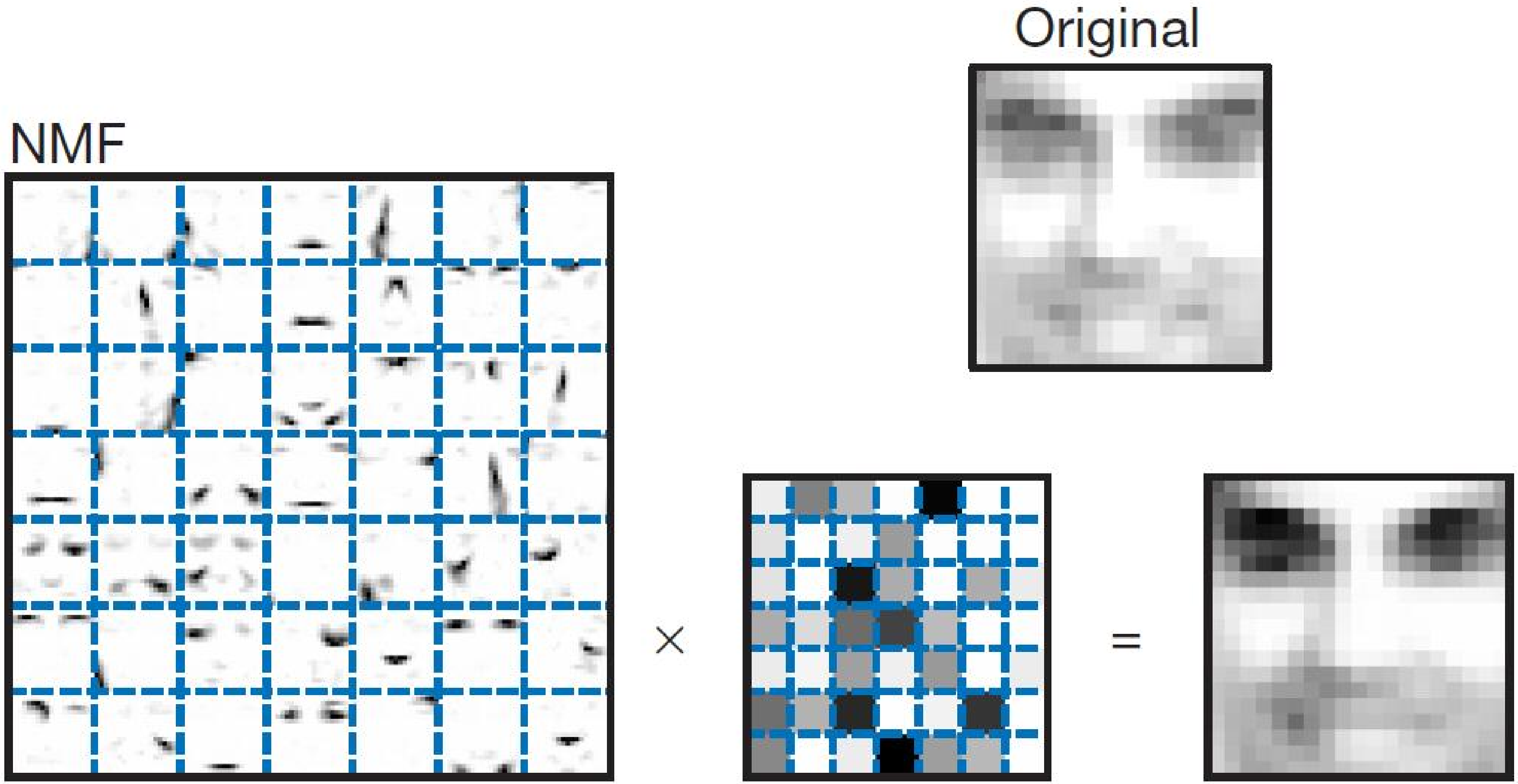
Vector quantization on face data (from Lee & Seung, Nature 1999)



PCA on face data



NMF of face data



NMF on complex polyphonic music

- NMF represents parts of the signal that fit the model (Virtanen, 2007)
- Individual drum instruments
- Repeating chords
- Any repetitive structure in the signal

Polyphonic example

- Original

- 20 separated components:

NMF algorithms

- NMF minimizes the error between **X** and **AS** while restricting **A** and **S** to be entry-wise non-negative
- Two commonly used distance measures (Lee & Seung 2001)
- Euclidean distance / L2 norm:

$$d_{euc} = \| \mathbf{X} - \mathbf{AS} \|_F^2$$

- Generalized Kullback-Leibler divergence:

$$d_{div}(X, AS) = \sum_{f,t} \mathbf{X}_{ft} \log(\mathbf{X}_{ft} / [\mathbf{AS}]_{ft}) - \mathbf{X}_{ft} + [\mathbf{AS}]_{ft}$$

- Many other measures

Multiplicative update rules

- Update rules which are guaranteed to be non-increasing
- Easy to implement and to extend
- Euclidean distance:

$$\mathbf{A} = \mathbf{A} \otimes \frac{\mathbf{X}\mathbf{S}^T}{(\mathbf{A}\mathbf{S})\mathbf{S}^T} \qquad \mathbf{S} = \mathbf{S} \otimes \frac{\mathbf{A}^T\mathbf{X}}{\mathbf{A}^T(\mathbf{A}\mathbf{S})}$$

- KL divergence

$$\mathbf{A} = \mathbf{A} \otimes \frac{(\mathbf{X}/(\mathbf{A}\mathbf{S}))\mathbf{S}^T}{\mathbf{1}\mathbf{S}^T} \qquad \mathbf{S} = \mathbf{S} \otimes \frac{\mathbf{A}^T(\mathbf{X}/\mathbf{A}\mathbf{S})}{\mathbf{A}^T\mathbf{1}}$$

where **1** is all-one matrix of size **X**

Optimization procedure

1. Initialize the entries in **A** and **S** with random positive values
2. Update **A**
3. Update **S**
4. Iterate steps 2 and 3

Also other optimization algorithms (e.g. projected steepest descent, Hoyer 2004)

NMF for audio in practice

- Calculate the magnitude spectrogram
  - Obtain each frame by multiplying the signal using a window function (for example 40 ms Hamming)
  - 50% or smaller frame shift
  - Calculate DFT in each frame *t*
  - Assign absolute values of the DFT to $\mathbf{X}_{ft}$
  - store the original phases
- Apply NMF (see previous slide) to obtain A and S
- Magnitude spectrogram of component k is obtained by
  - **A**(:,k) * **S**(k,:), or as **X**.*(**A**(:,k) * **S**(k,:)) ./ (**AS**) – Matlab notation
- Synthesis:
  - Assign the phases of the original mixture phase spectrogram to the separated component
  - Get time-domain frame by IDFT

NMF distance measures

- The distance measure should be chosen according to the properties of the data
- NMF can be viewed as maximum likelihood estimation
- Euclidean distance assumes additive Gaussian noise

$$p(\mathbf{X} \mid \mathbf{A}, \mathbf{S}) = \prod_{f,t} \mathrm{N}(\mathbf{X}_{f,t}; [\mathbf{AS}]_{f,t}, \sigma^2)$$

- KL assumes Poisson observation model (variance scales linearly with the model)

$$p(\mathbf{X} \mid \mathbf{A}, \mathbf{S}) = \prod_{f,t} \mathbf{Po}(\mathbf{X}_{f,t}; [\mathbf{AS}]_{f,t}) = \prod_{f,t} e^{-[\mathbf{AS}]_{f,t}} [\mathbf{AS}]_{f,t}^{\mathbf{X}_{ft}} / \mathbf{X}_{ft}!$$

- Equivalent to the multinomial model of PLSA

Bayesian approach (Virtanen and Cemgil 2008)

- Bayes rule: $p(\mathbf{A},\mathbf{S}|\mathbf{X}) = p(\mathbf{X}|\mathbf{A},\mathbf{S})\, p(\mathbf{A},\mathbf{S}) / p(\mathbf{X})$
- Allows us to place priors for **A** and **S**

  -> maximum a posterior estimation
- Typically sparse prior for the mixture weights
- Exponential prior $p(\mathbf{S}) = \prod_{k,t} \lambda e^{-\lambda \mathbf{S}_{kt}}$

  -> the objective to be minimized becomes (for example with the Gaussian model)

$$\| \mathbf{X} - \mathbf{AS} \| + \lambda \sum_{k,t} | \mathbf{S}_{kt} |$$

  -> non-negative sparse coding

Regularization in NMF

- Any cost terms can be added to the reconstruction error measure
  - Sparseness, temporal continuity (Virtanen 2007)
  - Correlation of weights (Wilson et al. 2008), spectra (Virtanen & Cemgil 2009)
  - Correlation of components (Wilson & Raj 2010)
- Optimization may become more difficult

Connection to PLSA

- Normalization not needed
- Slightly different probabilistic model formulation

Supervised NMF

- Prior information easy to include by training the spectral basis vectors in advance
- Source separation scenario:
  - Isolated training material of source 1 and source 2
  - Use NMF to train basis spectra for both sources separately
  - Combine the basis vector sets
  - Use NMF with the obtained basis vector set – keep the basis vectors fixed while updating the mixing weights
  - Synthesize source 1 by using its basis vectors only

Further analysis

- In practice a source source can be represented with more than one component
  - Cluster the components to sources
  - Supervised classification of components (train a classifier)
  - Example: separation of drums from polyphonic music by classification of NMF components by SVM (Helen & Virtanen 2005)
- Basis vectors are spectra
  - Pitch estimation (Vincent et al. 2007)
- Onset detection from mixture weights
  - Suits well for automatic drum transcription (Paulus & Virtanen 2005, Vincent et al. 2007)

Extensions of NMF

- Convolution in frequency
  – Translation of a basis vector in frequency: weight for each translation (Virtanen 2006)
  – With constant-Q spectral transformation allows modeling different pitches with a single basis vector
- Convolution in time
  – Basis vector extended to cover multiple adjacent frames -> time-varying spectra (Smaragdis 2007, Virtanen 2004)
  – Transpose of spectrogram -> equivalent to convolution in freq.
- Excitation-filter model (Heittola et al. 2009)
  – Each basis vector modeled as a sum of excitation and filter
- Harmonic bases (Vincent et al. 2007)
  – Each basis vector modeled as a weighted sum of harmonic combs with a limited frequency support

Voice separation demonstrations

- Demonstrations also available at http://www.cs.tut.fi/~tuomasv/

mixture	NMF-enhanced

mixture	sinusoidal model	binary mask	proposed

References

- Casey, M. and Westner, A., "Separation of Mixed Audio Sources by Independent Subspace Analysis", in *Proceedings of the International Computer Music Conference*, ICMA, Berlin, 2000.
- M. Plumbley, “Conditions for non-negative independent component analysis,” *IEEE Signal Processing Letters*, vol. 9, no. 6, pp. 177–180, 2002.
- K. W. Wilson and B. Raj, ”Spectrogram dimensionality reduction with independence constraints,” Int. Conf. on Audio, Speech, and Signal Processing, Dallas, USA, 2010, submitted for publication.
- D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *Adv. Neural Info. Proc. Syst.* 13, 556-562 (2001).
- D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788-791 (1999).
- T. Virtanen, *Monaural Sound Source Separation by Non-Negative Matrix Factorization with Temporal Continuity and Sparseness Criteria*, IEEE Transactions on Audio, Speech, and Language Processing, vol 15, no. 3, March 2007.
- P. O. Hoyer. “Non-negative Matrix Factorization with sparseness constraints” Journal of Machine Learning Research 5: 1457-1469, 2004.
- Helén, M., Virtanen, T., *Separation of Drums From Polyphonic Music Using Non-Negative Matrix Factorization and Support Vector Machine*, in proc. 13th European Signal Processing Conference Antalaya, Turkey, 2005.
- Paulus, J., Virtanen, T., *Drum Transcription with Non-negative Spectrogram Factorisation*, in proc. 13th European Signal Processing Conference Antalaya, Turkey, 2005

References (2)

- E. Vincent, N. Bertin, R. Badeau “Two Nonnegative matrix factorization methods for polyphonic pitch transcription”. *Proc. of the International Conf. on Music Information Retrieval (ISMIR)*, Vienne, 2007.
- T. Virtanen, A. T. Cemgil, and S. J. Godsill. *Bayesian Extensions to Non-negative Matrix Factorisation for Audio Signal Modelling*, ICASSP 2008 .
- Wilson, K.W., B. Raj, and P. Smaragdis, 2008. Regularized Non-Negative Matrix Factorization with Temporal Dependencies for Speech Denoising. In proceedings of Interspeech 2008, Brisbane, Australia, September 2008.
- T. Virtanen and A. T. Cemgil. *Mixtures of Gamma Priors for Non-Negative Matrix Factorization Based Speech Separation*, in Proc. ICA 2009, Paraty, Brazil,2009.
- T. Virtanen. ”Sound Source Separation in Monaural Music Signals”, PhD Thesis, Tampere University of Technology, 2006.
- T. Virtanen, *Separation of Sound Sources by Convolutive Sparse Coding*, ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing, SAPA 2004.
- T. Heittola, A. Klapuri, and T. Virtanen. *Musical Instrument Recognition in Polyphonic Audio Using Source-Filter Model for Sound Separation*, to be presented in Proc. 10th Int. Society for Music Information Retrieval Conf. (ISMIR 2009), Kobe, Japan, 2009.
- Smaragdis, P. 2007. Convolutive Speech Bases and their Application to Speech Separation. In IEEE Transactions of Speech and Audio Processing. January 2007
- D. Ellis and D.F Rosenthal (1998) Mid-level representations for Computational Auditory Scene Analysis, Chapter 17 in *Computational auditory scene analysis*, D. F. Rosenthal and H. Okuno, eds., Lawrence Erlbaum, pp. 257-272, 1998.