Compositional models for sound source separation and processing

Tuomas Virtanen

Tampere University of Technology
Laboratory of Signal Processing
Finland
www.cs.tut.fi/~tuomasv/

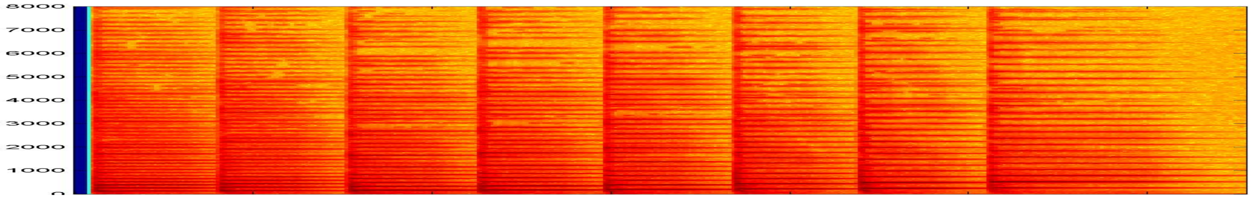
TAMPERE UNIVERSITY OF TECHNOLOGY

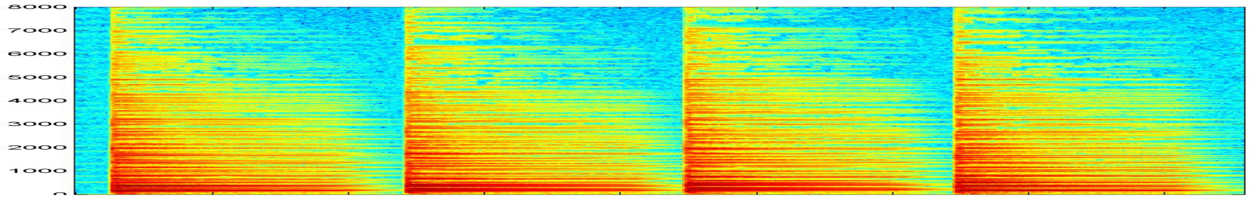
Outline

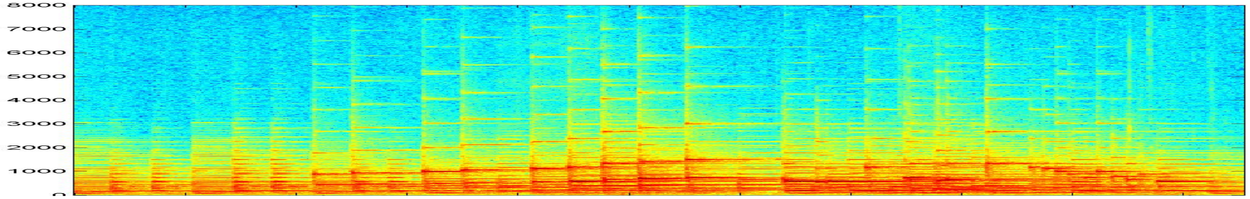
- Introduction
  - What are compositional models
  - Signal representation
- Application: source separation
- Algorithm: non-negative matrix factorization (NMF)
- Analyzing the semantics of sound
- Model alternatives
- Comparison to DNNs
- Missing data techniques

Compositional model

- In general, sounds do not cancel each other
- Typically, individual components combine to form the sounds we hear
  - Notes, but also multiple instruments, form music

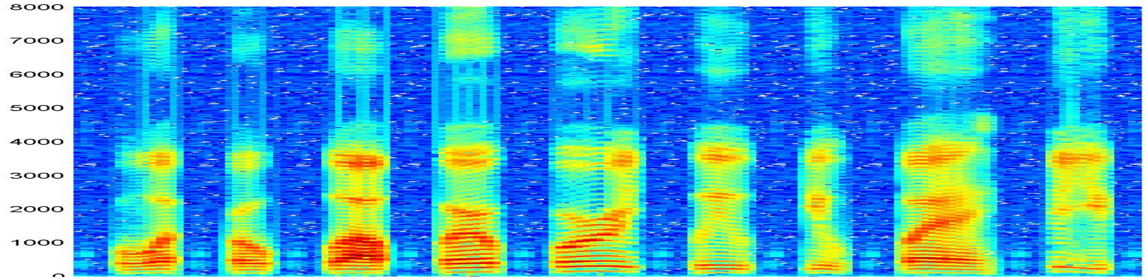


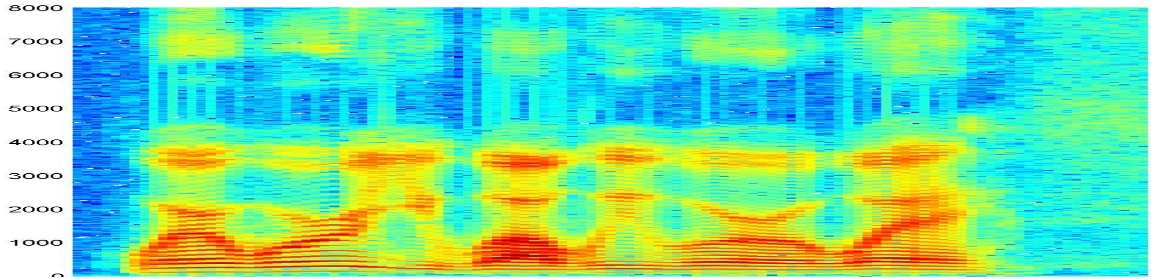




Compositional model

- In general, sounds do not cancel each other
- Typically, individual components combine to form the sounds we hear
  - Notes, but also multiple instruments, form music
  - Phoneme-like sounds combine to form speech





Compositional model

- In general, sounds do not cancel each other
- Typically, individual components combine to form the sounds we hear
  - Notes, but also multiple instruments, form music
  - Phoneme-like sounds combine to form speech
- The *compositional model* is a linear, additive combination of components that do not result in subtraction or diminishment of any of the constituents

Compositional model

- Feature vector $\boldsymbol{y}_t$ is decomposed into weighted sum of basis vectors $\boldsymbol{a}_n$

$$\boldsymbol{y}_t \approx \sum_n \boldsymbol{a}_n x_{nt}$$

- x_{nt} are gains of the components in observation t
- Compositional model: both the basis vectors and weights are constrained to be <u>**non-negative**</u>

Compositional model

- Model in a vector-matrix form

$$\begin{bmatrix} y_{1t} \\ \vdots \\ y_{Ft} \end{bmatrix} \approx \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{F1} & \cdots & a_{FN} \end{bmatrix} \cdot \begin{bmatrix} x_{1t} \\ \vdots \\ x_{Nt} \end{bmatrix} \Rightarrow \boldsymbol{y}_t \approx \boldsymbol{A}\boldsymbol{x}_t$$

Model for multiple observations

- We can efficiently write the compositional model

$$\boldsymbol{y}_t \approx \boldsymbol{A}\boldsymbol{x}_t, \qquad t = 1 \dots T$$

- for all T observations (e.g. a spectrogram), as:

$$[\boldsymbol{y}_1 \dots \boldsymbol{y}_T] \approx \boldsymbol{A} \cdot [\boldsymbol{x}_1 \dots \boldsymbol{x}_T]$$

- Or even:

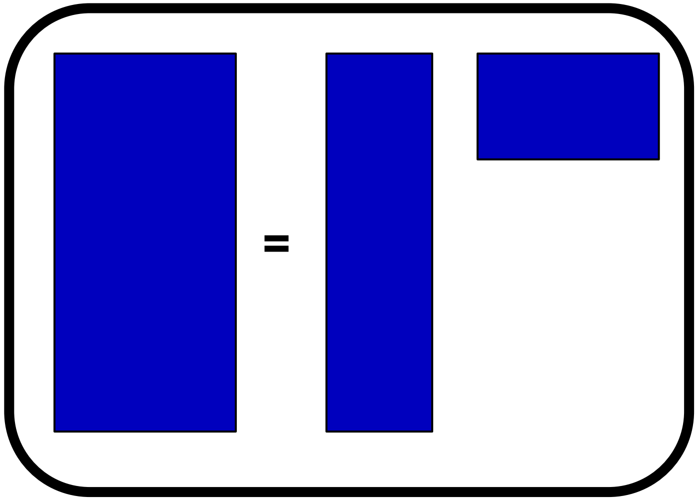
$$\boldsymbol{Y} \approx \boldsymbol{A}\boldsymbol{X}$$

Visual representation

$$\boldsymbol{Y} \approx \boldsymbol{A}\boldsymbol{X}$$

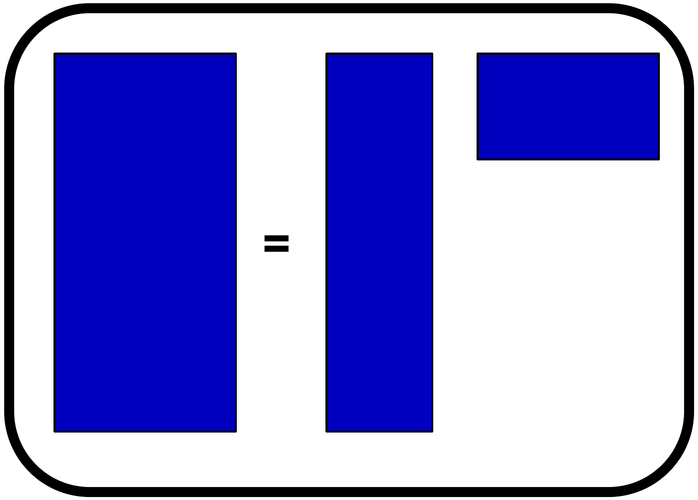
Visual representation

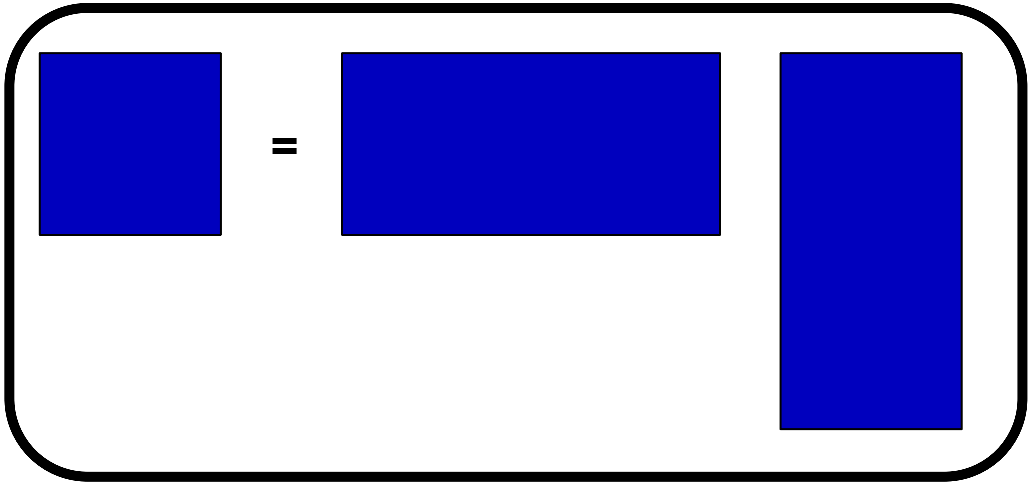
$$Y \approx AX$$



Visual representation

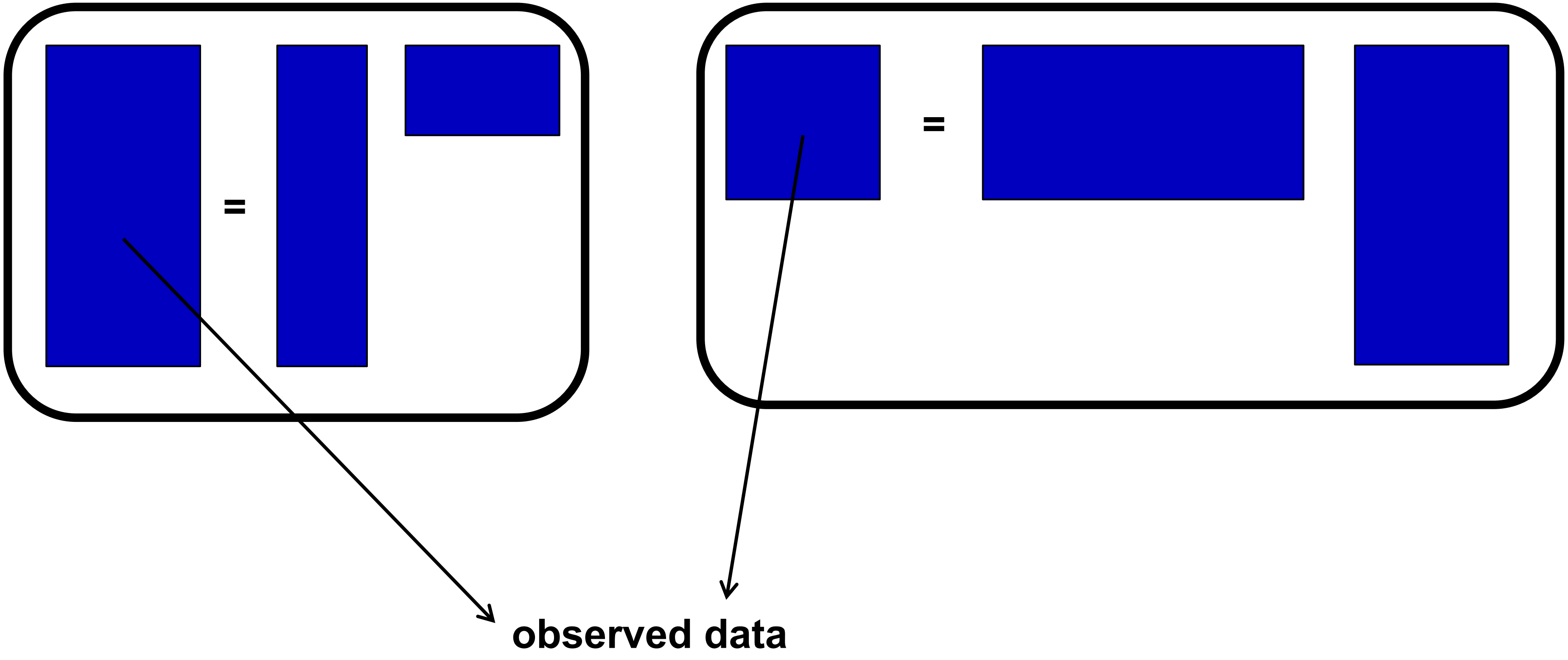
$$Y \approx AX$$





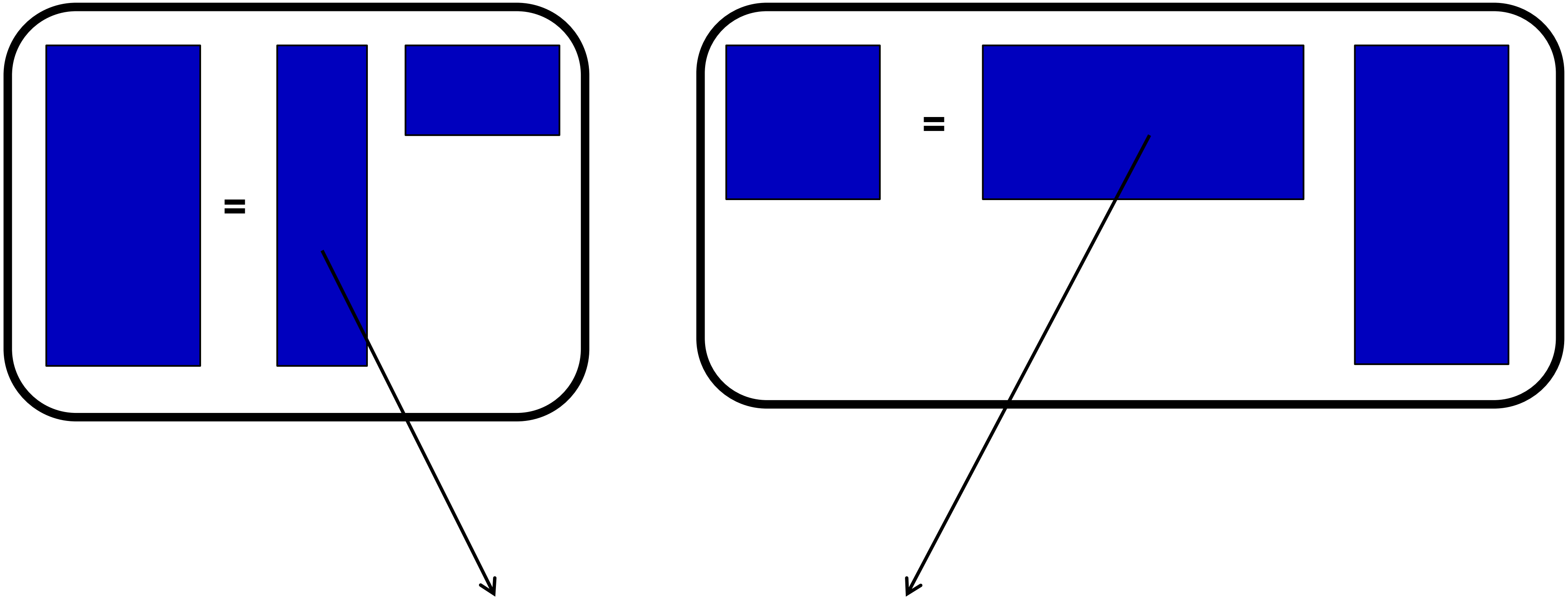
Visual representation

$$Y \approx AX$$



Visual representation

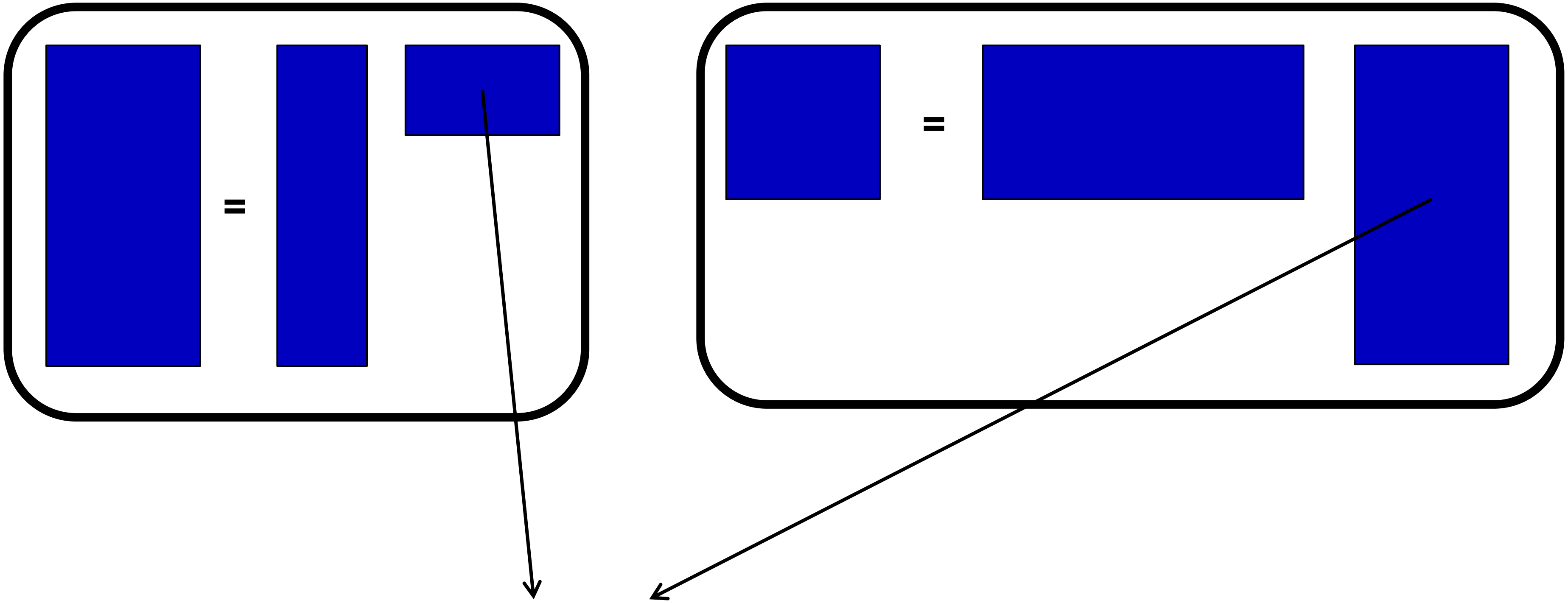
$$Y \approx AX$$



dictionary **or** ***basis*** **(learned or constructed, updated or kept fixed)**

Visual representation

$$Y \approx AX$$



mixture weights, sparse representations

Compositional generative model

- The compositional model explains how the observations are generated, given the model parameters

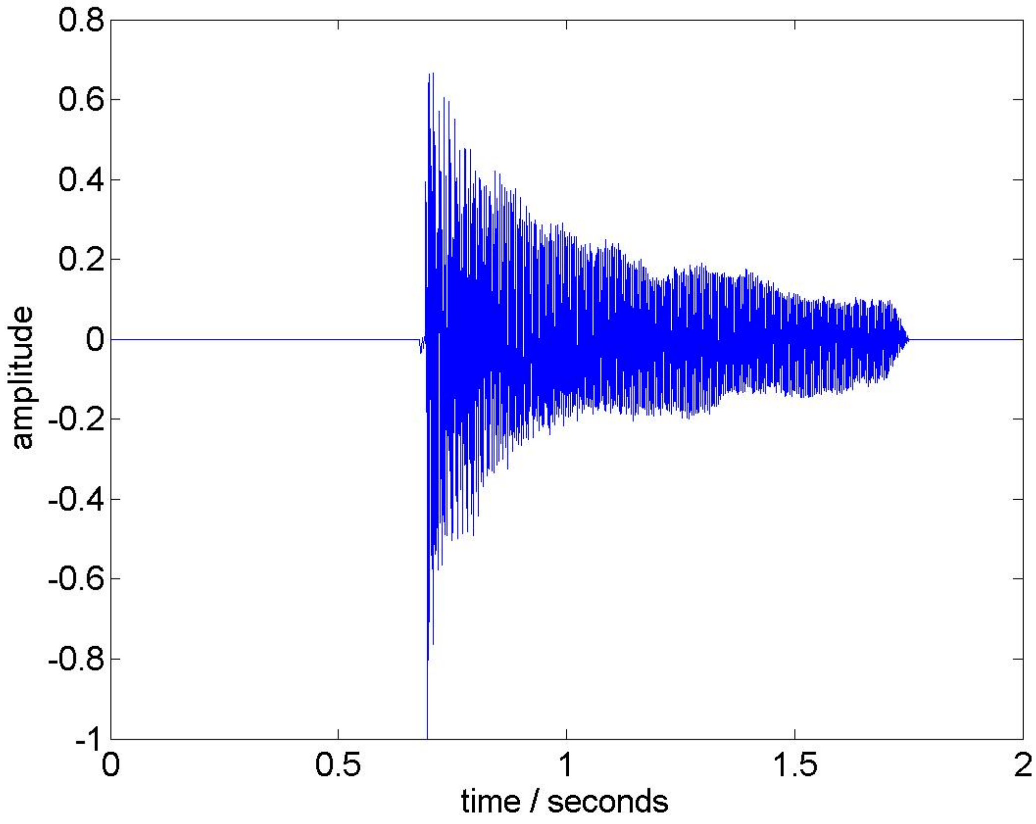
$$\boldsymbol{p}(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{X})$$

Audio representation

- Compositional models require a non-negative representation
- Audio *signals* have both negative and positive values
- Need for a mid-level representation that is used for processing

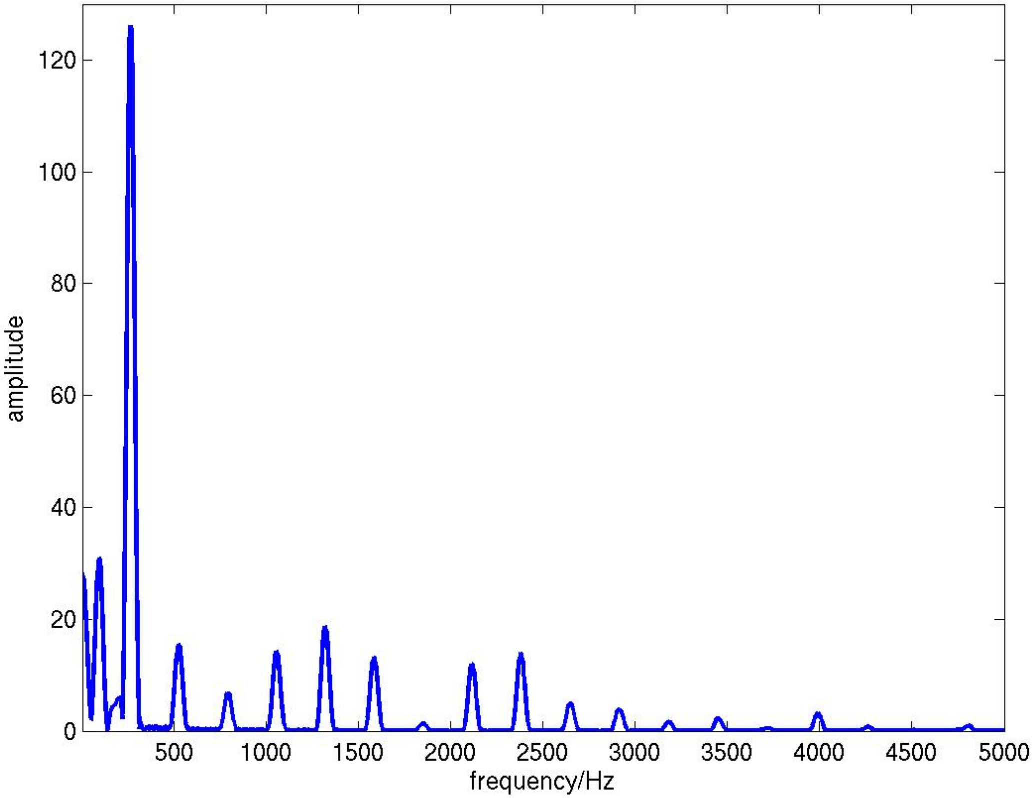
Audio representation

- Audio signal – amplitude as a function of time



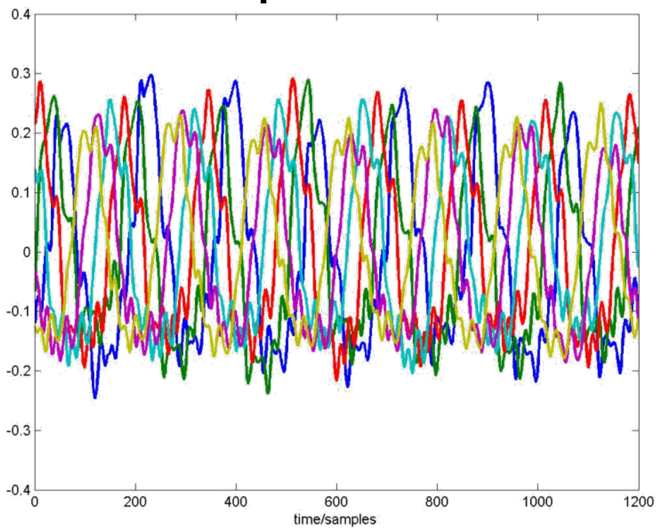
Magnitude spectrum

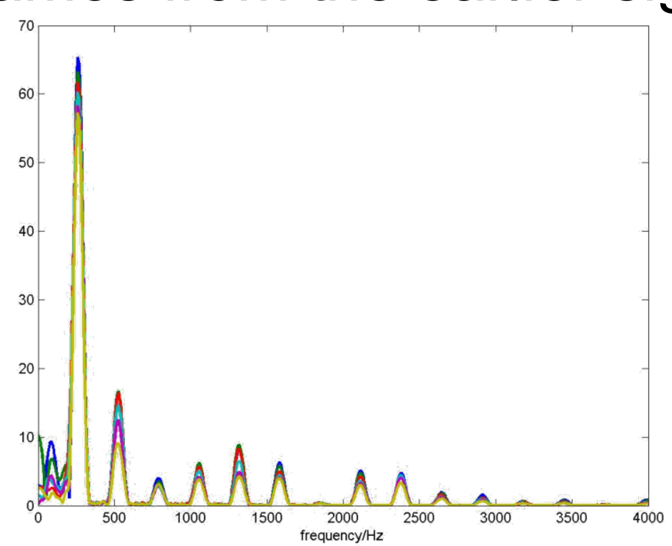
- Phases are discarded and only the magnitudes are used → non-negative representation
- Can use any spectral resolution (linear, logarithmic, perceptual…)



Magnitude spectrum

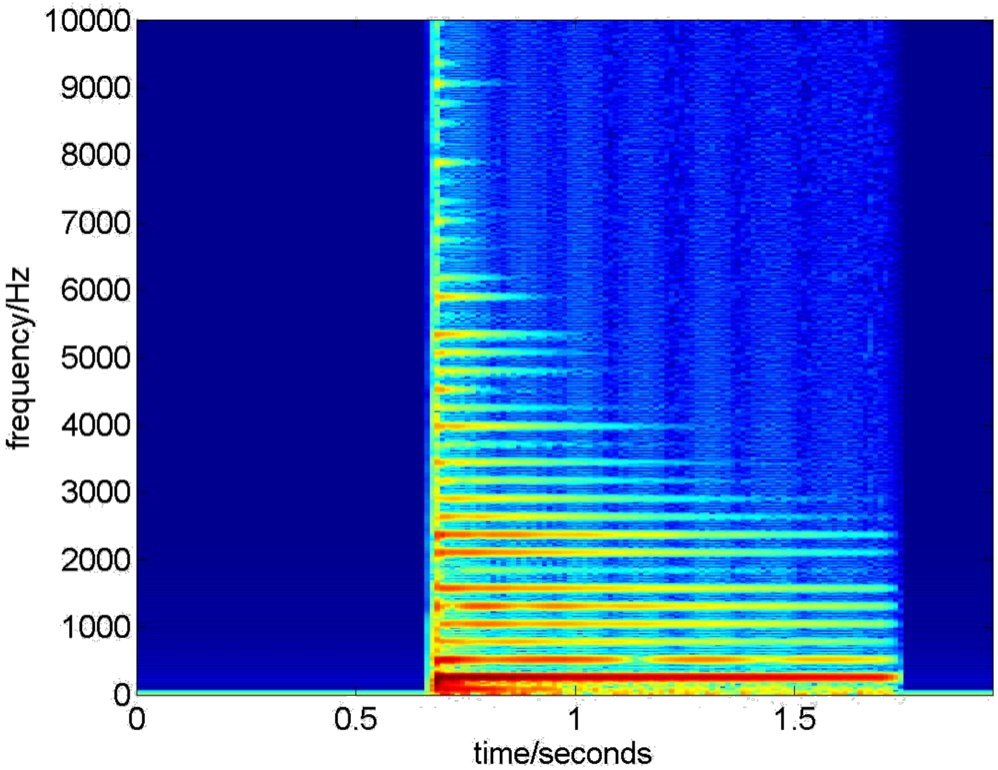
- Natural sounds have a clear structure in the magnitude spectrum domain
- Discarding the phases makes the representation invariant to many factors
  - Relative window position
  - Phase of the acoustic impulse response from source to microphone
- Example: five consecutive frames from the earlier signal





Spectrogram

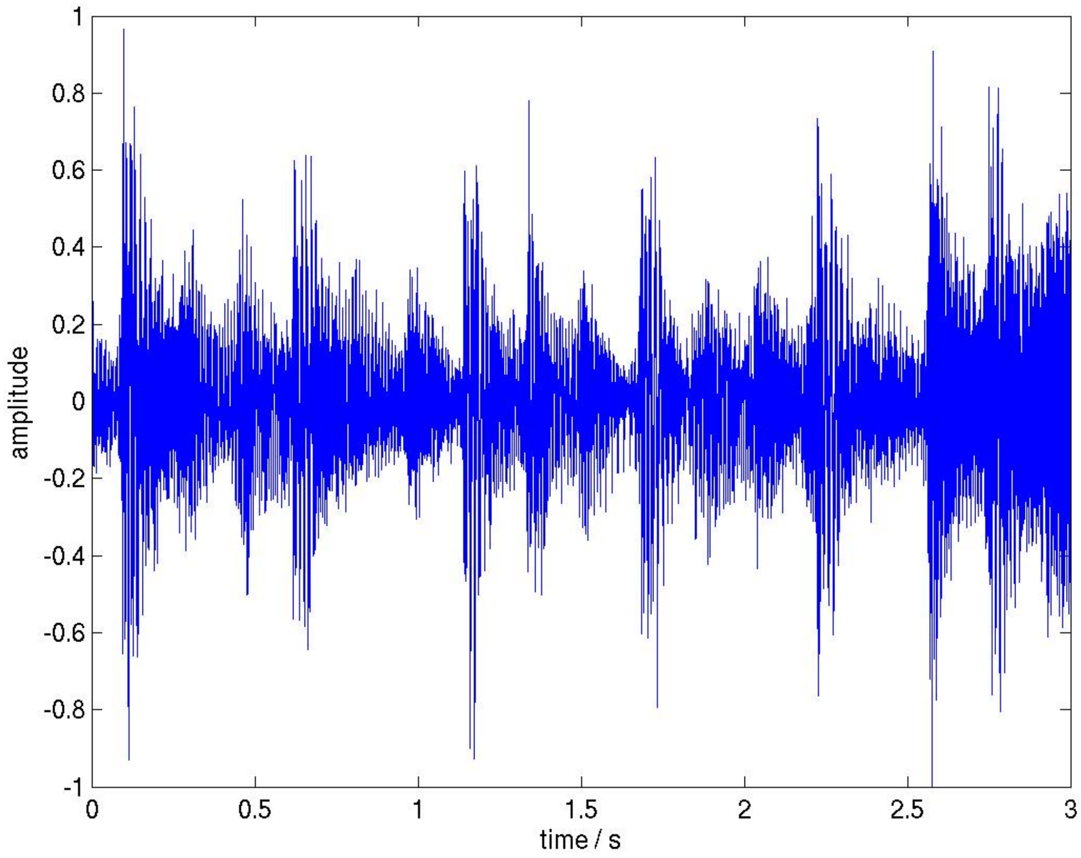
- Spectra in each frame grouped to a matrix
- Represents the intensity of a sound as a function of time and frequency



Linear superposition

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

$$y(t) = s_1(t) + s_2(t)...$$



Additivity of magnitude spectra

- In the magnitude spectral domain, sounds are *approximately* additive

$$y(t) = s_1(t) + s_2(t)...$$

$$|Y(f)| \approx |S_1(f)| + |S_2(f)|...$$

- Exactly additive only when the phases are coherent
- For independent source, power spectra are additive inthe expectation sense:

$$E\left\{|Y(f)|^2\right\} = E\left\{|S_1(f)|^2\right\} + E\left\{|S_2(f)|^2\right\}...$$

- … sounds are also approximately additive in the power spectral domain;

$$|Y(f)|^2 \approx |S_1(f)|^2 + |S_2(f)|^2 ...$$

Additivity of magnitude spectra

- Magnitude vs. power spectrum representation?
  - I.e., $|Y(f)|$ vs. $|Y(f)|^2$
- Determines the dynamic scale of the representation
- Affects the relative importance of low vs. high-intensity observations
- Related to the compositional model estimation criterion (see later)
- Empirically observed that additivityof magnitudes works better

B. King, C. Fevotte, and P. Smaragdis, "Optimal cost function and magnitude power for NMF-based speech separation and music interpolation," in *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing*, Santander, Spain, 2012.

Additivity of magnitude spectra

- How valid is the approximation?
- Natural sounds are *sparse* and therefore *disjoint* in the time-frequency domain

$$|S_1(t,f)||S_2(t,f)| \approx 0$$

- Additivity of magnitude or power spectrum works well enough in practice
- Lower frequency and time resolutions lead to lower sparseness and disjointness

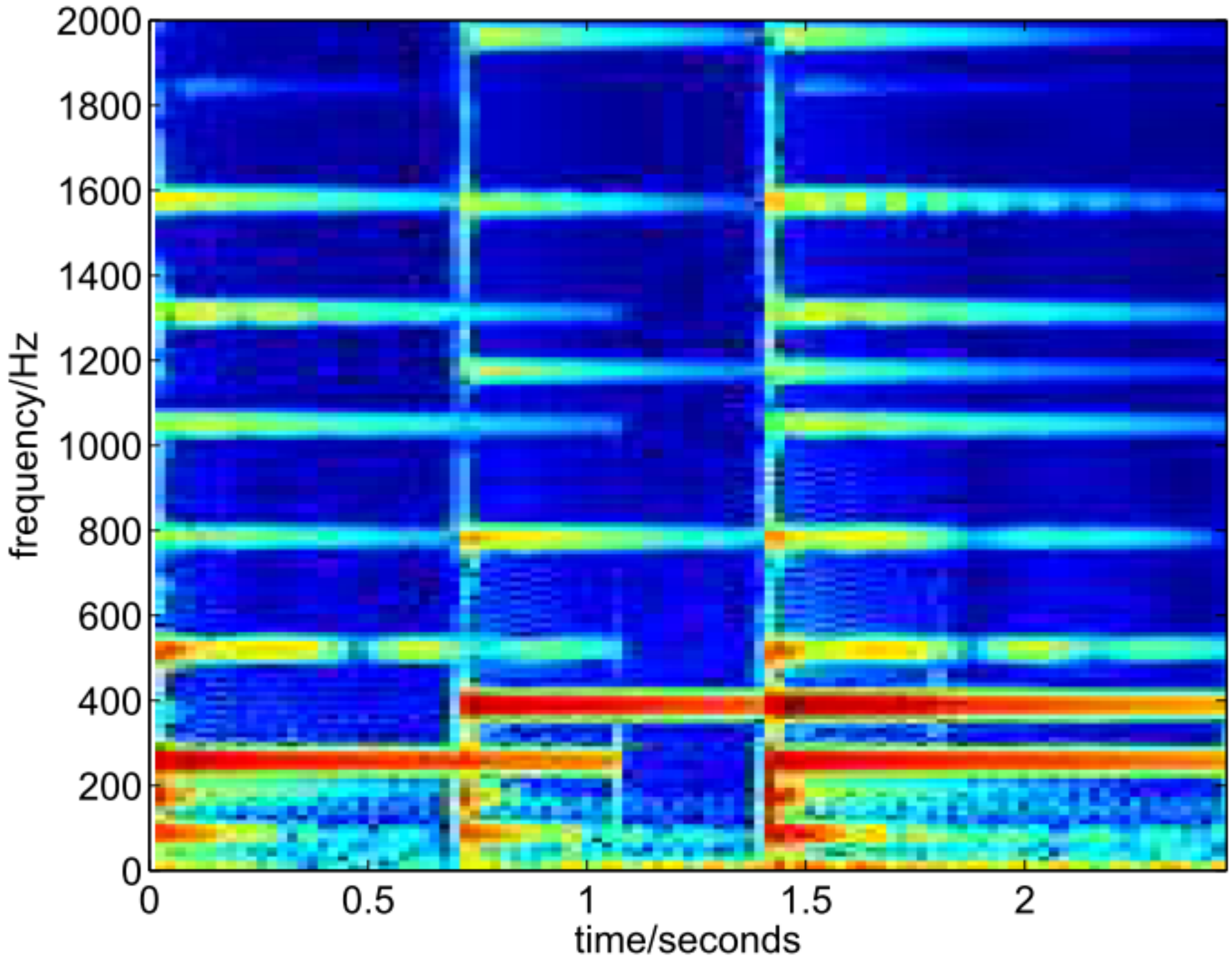
Compositional spectral model

- Clear interpretation:
  - Signal modeled as a sum of components
  - Each components has a fixed spectrum (basis vectors $\boldsymbol{a}_n$) and time-varying gain x_{nt}

$$\boldsymbol{y}_t \approx \sum_n \boldsymbol{a}_n x_{nt}$$

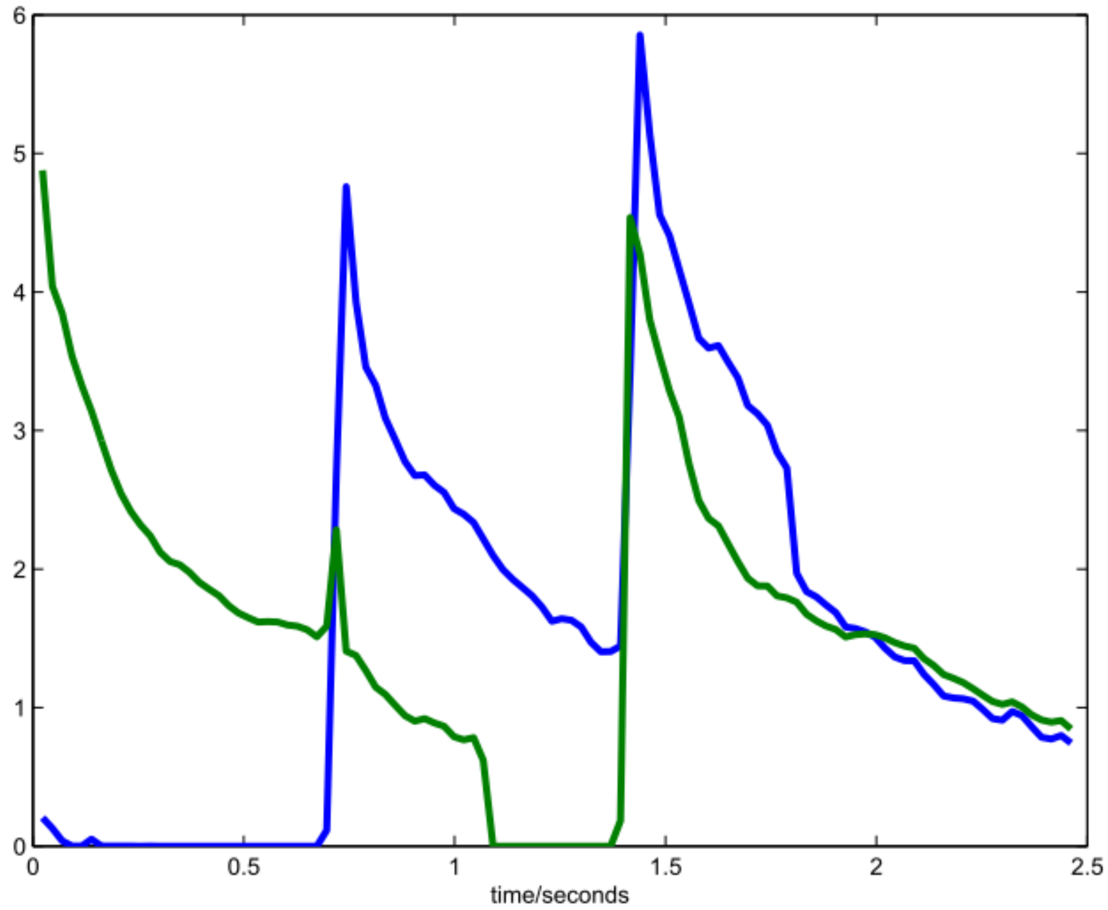
T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066 - 1074, 2007.

Mixture spectrogram



Results with NMF

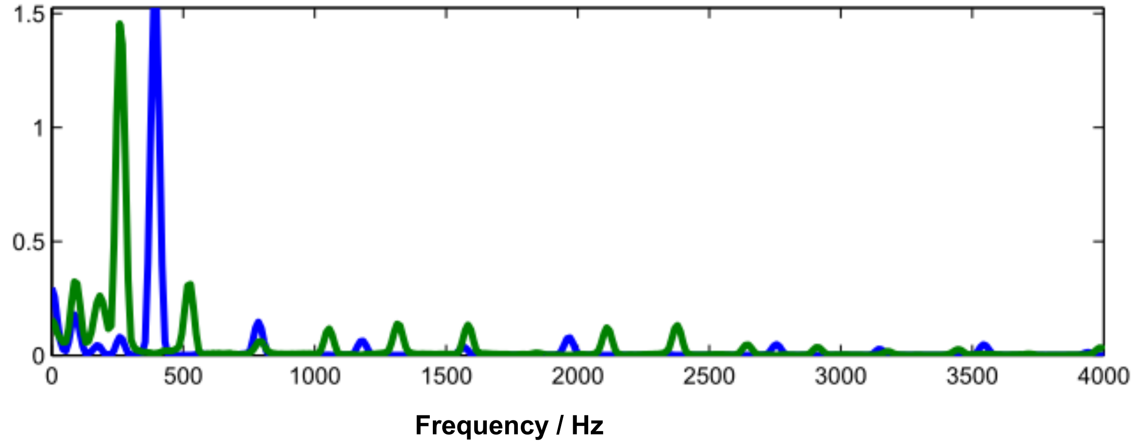
- Weights over time: separation of notes



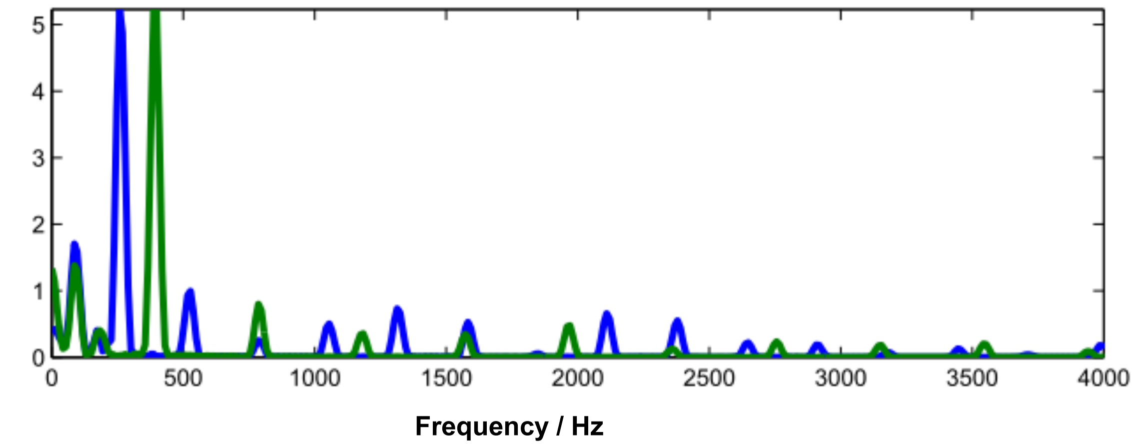
Results with NMF

- Bases correspond to individual notes

NMF

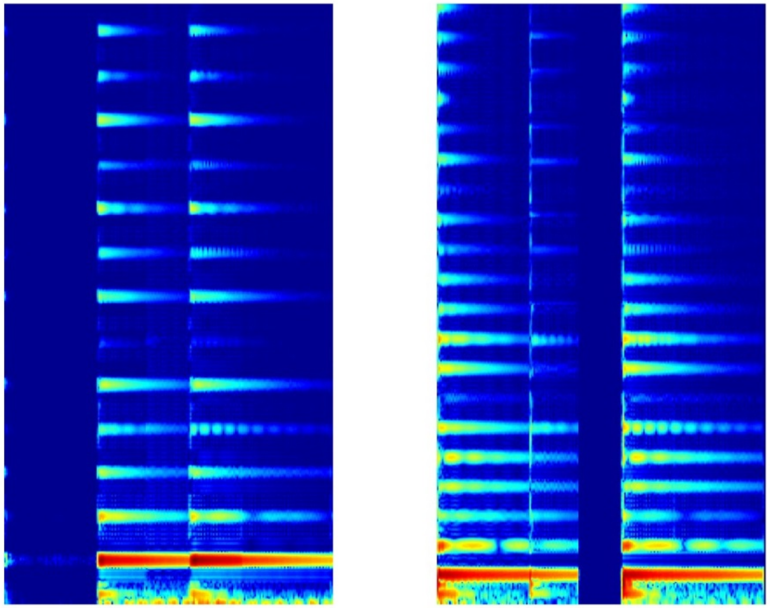


Original



Wiener-style reconstruction

- *k*-th component reconstructed as: $\mathbf{Y} \otimes \frac{\mathbf{A}(:,k) \bullet \mathbf{X}(k,:)}{\mathbf{AX}}$



Signal reconstruction

- For each component:
  1. Use the phases of the mixture signal
  2. IFFT
  3. Overlap-add

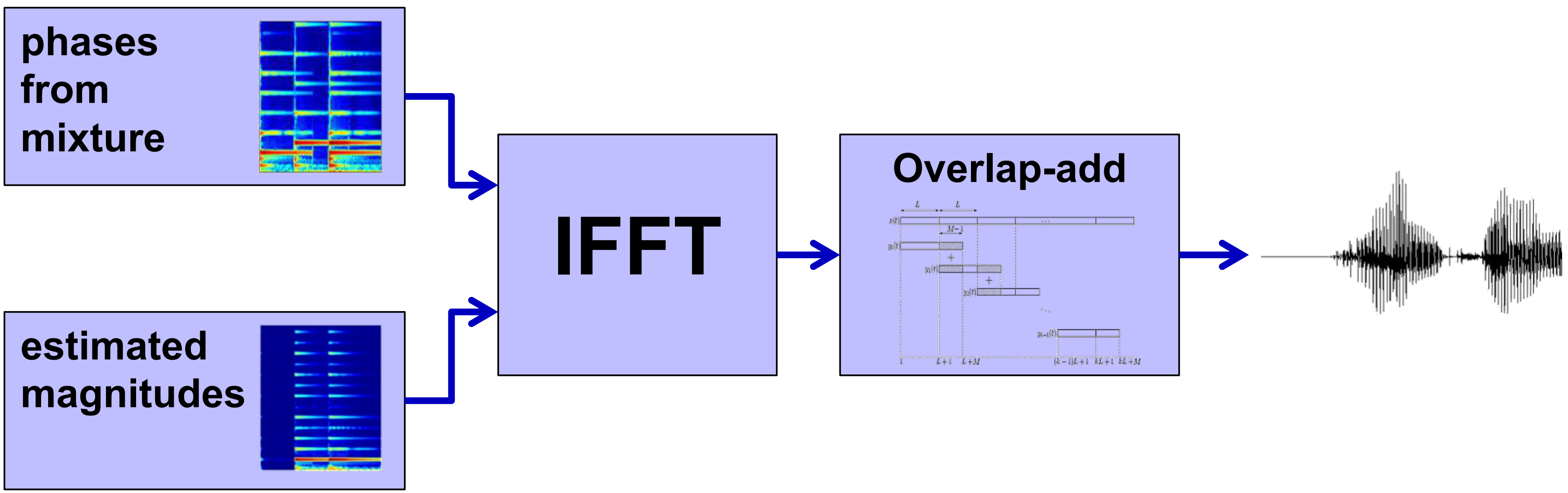


Figure source: Wikipedia Commons, *Paolo Serena, University of Parma (Italy)*

Non-negative matrix factorization

- Minimize error $d(\mathbf{Y}, \mathbf{AX})$ between **Y** and **AX** while restricting **A** and **X** to be entry-wise non-negative

$$\mathbf{A}^{*}, \mathbf{X}^{*} = \arg\min_{\mathbf{A},\mathbf{X}} d(\mathbf{Y}, \mathbf{AX})$$

- Supervised NMF – estimate only the weights, the bases are given:

$$\mathbf{X}^{*} = \arg\min_{\mathbf{X}} d(\mathbf{Y}, \mathbf{AX})$$

NMF criteria

- Different distance measures
- Squared error (L2 norm):

$$d_{\mathrm{SQ}}(\mathbf{Y},\mathbf{AX}) = \sum_{f,t} (\mathbf{Y}_{ft} - [\mathbf{AX}]_{ft})^2 = \| \mathbf{Y} - \mathbf{AX} \|_F^2$$

- Generalized Kullback-Leibler divergence:

$$d_{\mathrm{KL}}(\mathbf{Y},\mathbf{AX}) = \sum_{f,t} \mathbf{Y}_{ft} \log(\mathbf{Y}_{ft} / [\mathbf{AX}]_{ft}) - \mathbf{Y}_{ft} + [\mathbf{AX}]_{ft}$$

- Itakura-Saito divergence

$$d_{\mathrm{IS}}(\mathbf{Y},\mathbf{AX}) = \sum_{f,t} \mathbf{Y}_{ft} / [\mathbf{AX}]_{ft} - \log(\mathbf{Y}_{ft} / [\mathbf{AX}]_{ft})$$

- Each of these correspond to different generative model $\boldsymbol{p}(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{X})$

- D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of Neural Information Processing Systems*, Denver, USA, 2000, pp. 556-562.
C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793-830, 2009.

NMF algorithms

- The objective function is biconvex

$$\mathbf{A}^*, \mathbf{X}^* = \arg\min_{\mathbf{A},\mathbf{X}} d(\mathbf{Y}, \mathbf{A}\mathbf{X})$$

- Global optimum cannot be found
- Iterative algorithms which repeatedly update **A** and **X** so that the cost decreases at each iteration

Multiplicative update rules

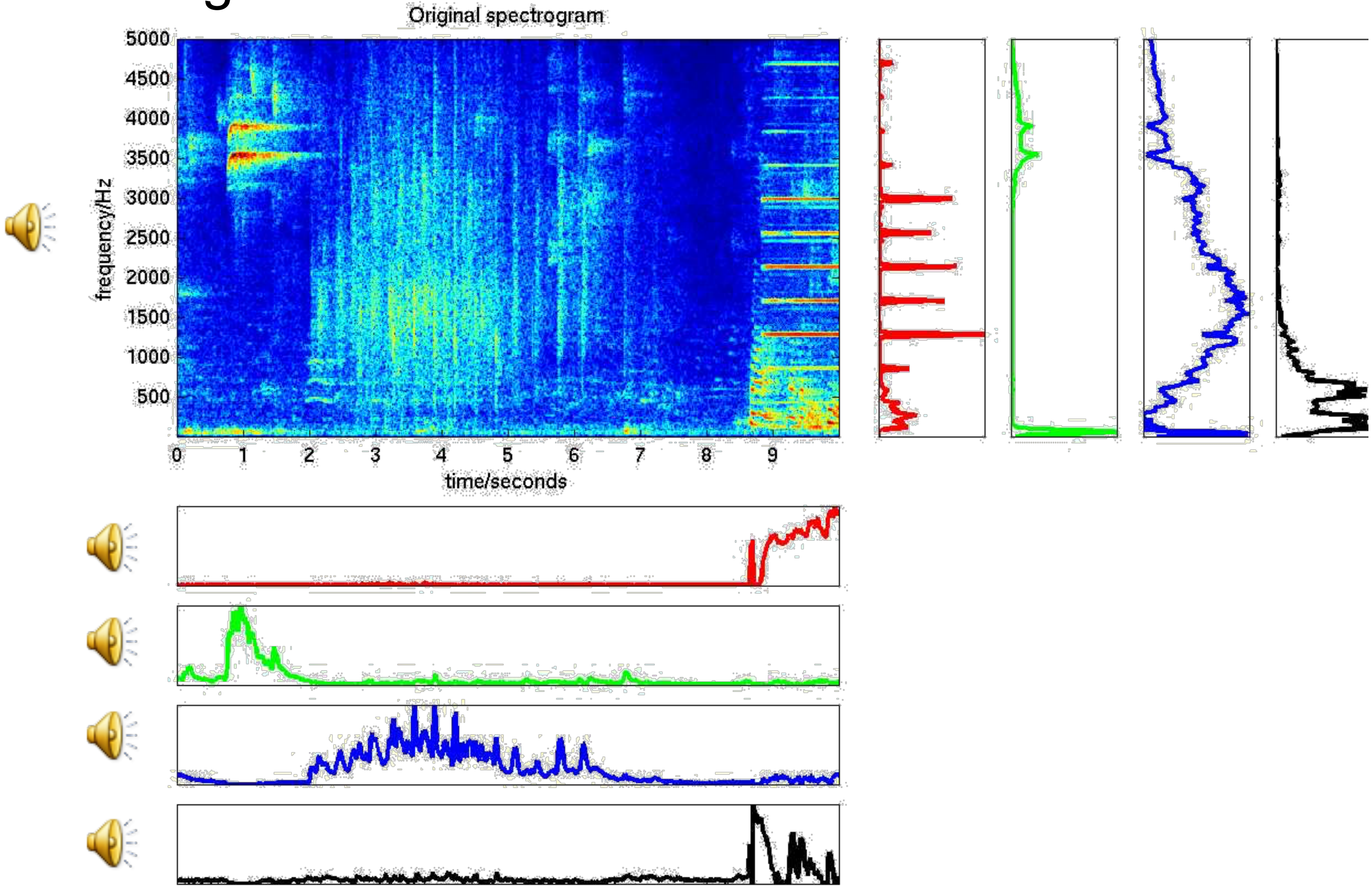
- Update rules under which the cost are guaranteed to be non-increasing
- Guarantees non-negativity of the parameters
- Easy to implement and to extend
- Updates for the KL divergence

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\mathbf{A}^{\mathrm{T}}(\mathbf{Y}/\mathbf{AX})}{\mathbf{A}^{\mathrm{T}}\mathbf{1}} \qquad \mathbf{A} \leftarrow \mathbf{A} \otimes \frac{(\mathbf{Y}/\mathbf{AX})\mathbf{X}^{\mathrm{T}}}{\mathbf{1}\mathbf{X}^{\mathrm{T}}}$$

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

Real-world examples

- Basketball game



Real-world examples

- High-quality separation of complex auditory scenes in blind manner not achievable
- Multiple components required to represent an invidual source
- Each component still corresponds to semantically meaningful entity

Supervised source separation

- Prior information easy to include by training the spectral basis **A** vectors in advance
- Optimization problem is convex and therefore finding the global optimum is guaranteed
- More efficient algorithms

$$\mathbf{X}^{|*} = \arg\min_{\mathbf{X}} d(\mathbf{Y}, \mathbf{A}\mathbf{X})$$

- Yields impressive results in matched conditions

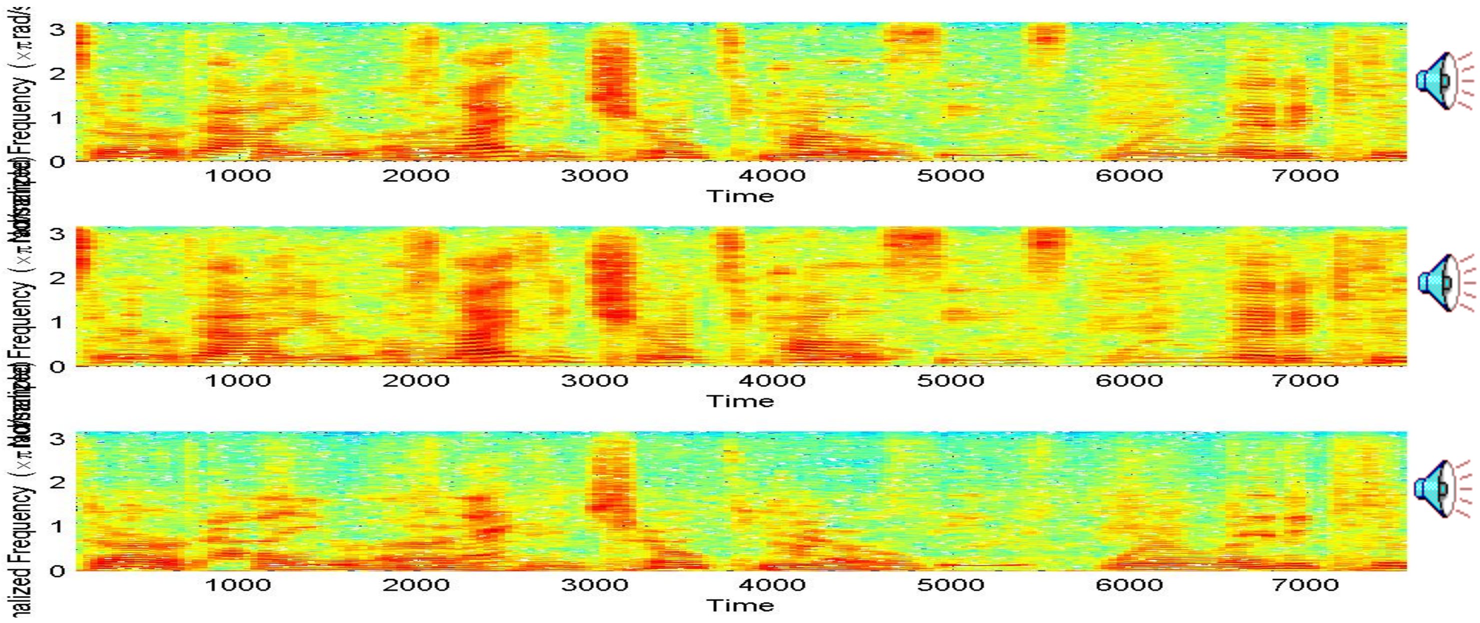
Supervised source separation

- Source separation (SS) scenario:
  - Isolated training material of speech (s) and noise (n)
  - Obtain basis spectra for each source separately
  - Concatenate the dictionaries:

$$\mathbf{Y} = \mathbf{S} + \mathbf{N} \approx \hat{\mathbf{S}} + \hat{\mathbf{N}} = \mathbf{A}_s \mathbf{X}_s + \mathbf{A}_n \mathbf{X}_n = \mathbf{A}\mathbf{X}$$

  - Use NMF with the obtained dictionary – keep the dictionary fixed while updating the mixing weights
  - Synthesize each source by using only its own basis vectors

Separate overlapping speech



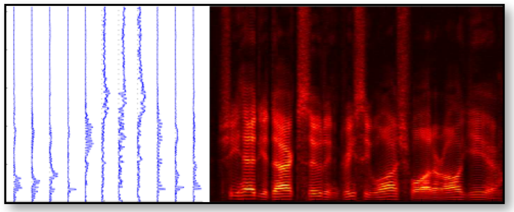
- Bases for both speakers learnt from 5 second recordings of individual speakers

Semi-supervised separation

- We may not have training data for all sources
  - But we usually know some

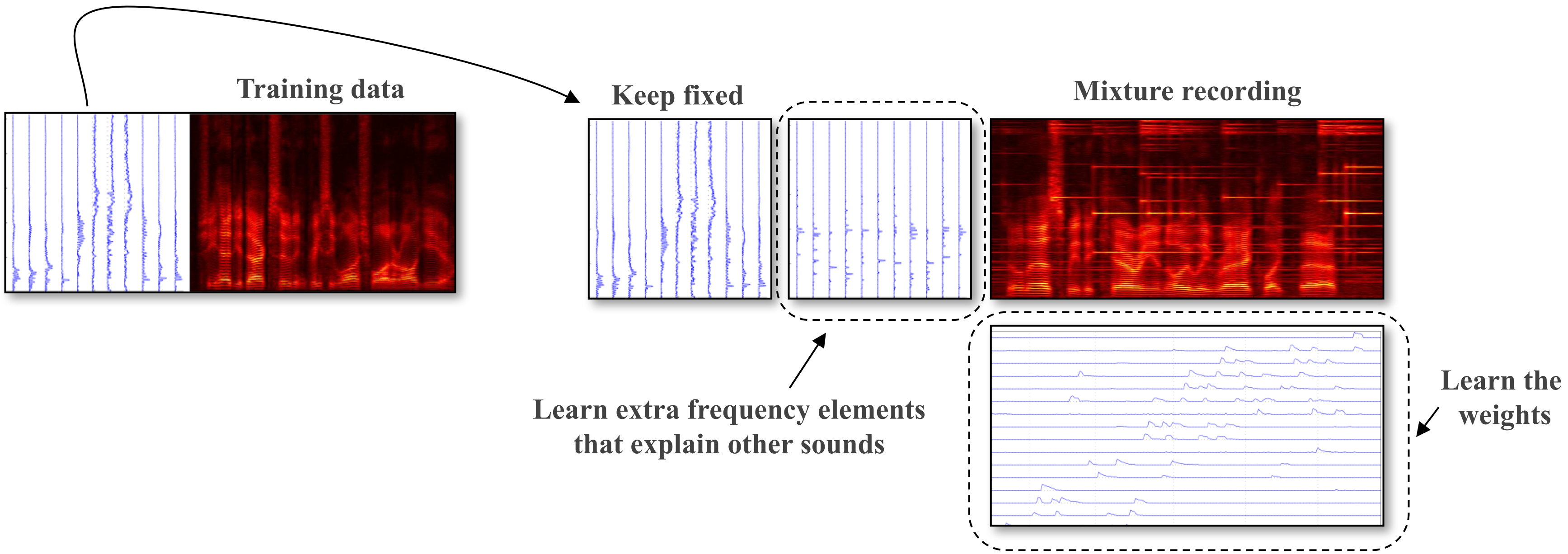
Semi-supervised separation

- We may not have training data for all sources
  - But we usually know some
- Two steps:
  - Supervised: Train dictionary for known sources



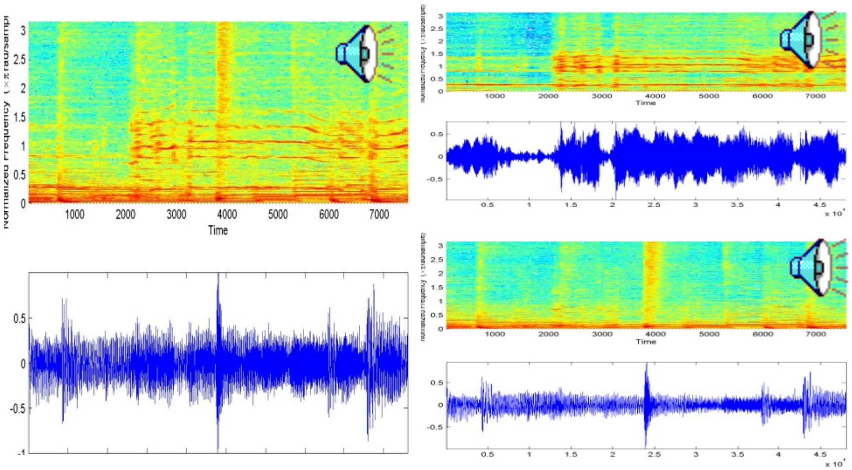
Semi-supervised separation

- We may not have training data for all sources
  - But we usually know some
- Two steps:
  - Supervised: Train dictionary for known sources
  - Unsupervised: Train *part* of the dictionary unsupervised on target mixture

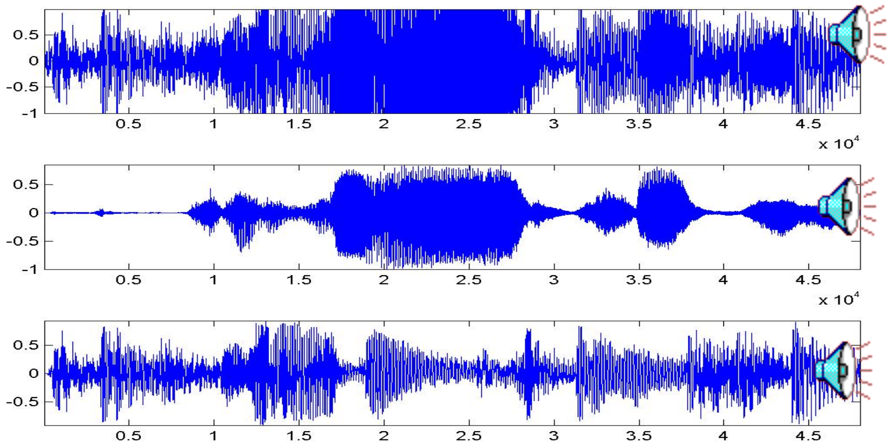


Smaragdis, P. Raj, B. and Shashanka, M.V. 2007. Supervised and Semi-Supervised Separation of Sounds from Single-Channel Mixtures. In proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation. London, UK. September 2007.

Semi-supervised separation



“Raise my rent” by David Gilmour

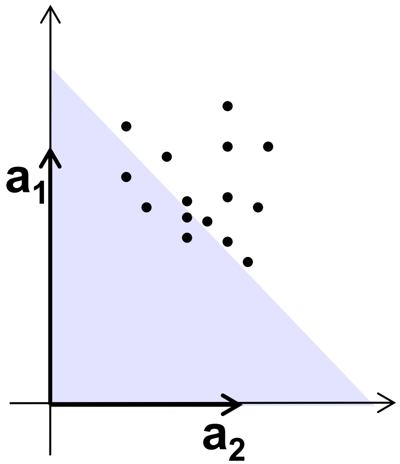


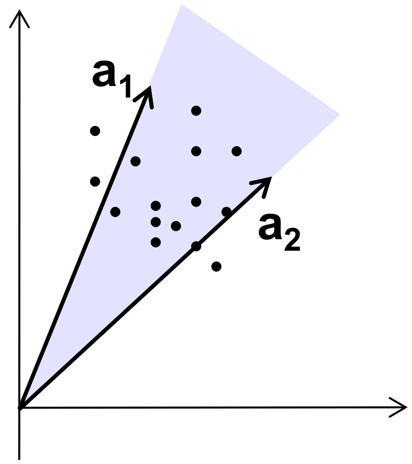
Norah Jones singing “Sunrise”

- Background music “bases” learnt from 5-seconds of music-only segments within the song
- Guitar bases/voice learnt from the rest of the song

What is a good dictionary?

- It should be kept relatively small
  - Overcomplete dictionaries have non-unique solutions without regularization
  - To reduce computational complexity
- It should be capable of accurately describing the source, and generalize well to unseen data
- It should be discriminative: sources cannot be well represented using a dictionary of another source





What is a good dictionary?

- It should be kept relatively small
  - Overcomplete dictionaries have non-unique solutions without regularization
  - To reduce computational complexity
- It should be capable of accurately describing the source, and generalize well to unseen data
- It should be discriminative: sources cannot be well represented using a dictionary of another source
  - In the field of *sparse representations*, this is stated as: a source should be sparse in one dictionary and dense in the other

Dictionary learning

- Non-negative matrix factorization (NMF)
- Clustering
  - k-means clustering
  - Hierarchical clustering
- Approaches used in the field *sparse representations* and *compressed sensing* (CS)
  - Attempt to find dictionaries which *sparsely* represent sources
  - Generally no non-negativity constraints
  - Some exceptions, e.g. non-negative K-SVD

- **M. Aharon, M. Elad, and A. Bruckstein, "K-SVD and its non-negative variant for dictionary design," in Proceedings of SPIE Conference on Wavelet Applications in Signal and Image Processing XI, San Diego, USA, 2005.**
- **R. G. Baraniuk, "Compressive sensing," IEEE Signal Processing Magazine, vol. 24, no. 4, pp. 118-121, 2007.**

Dictionary learning

- Non-negative matrix factorization (NMF)
- Clustering
  - k-means clustering
  - Hierarchical clustering
- Approaches used in the field *sparse representations* and *compressed sensing* (CS)
  - Attempt to find dictionaries which *sparsely* represent sources
  - Generally no non-negativity constraints
  - Some exceptions, e.g. non-negative K-SVD
- Pros and cons
  - 👍 Dictionaries generalize well to unseen data
  - 👍 NMF and CS approaches consider additivity: smaller, parts-based dictionaries
  - 👎 Parts-based representations are often less discriminative between sources

Dictionary sampling

- Approach: directly use samples from the training data
  - Often called “exemplars”
  - This may lead to very large dictionaries!

- P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in Proceedings of Neural Information Processing Systems, Vancouver, Canada, 2009.
- J. F. Gemmeke, H. Van hamme, B. Cranen, and L. Boves, “Compressive sensing for missing data imputation in noise robust speech recognition,” IEEE Journal of Selected Topics in Signal Processing, vol. 4, no. 2, pp. 272–287, 2010.

Dictionary sampling

- Approach: directly use samples from the training data
  - Often called “exemplars”
  - This may lead to very large dictionaries!
- Common techniques:
  - random sampling: a random subset of the exemplars
  - pruning: select a subset using some criterion
    - Correlation between exemplars
    - How often an exemplar is activated on development data

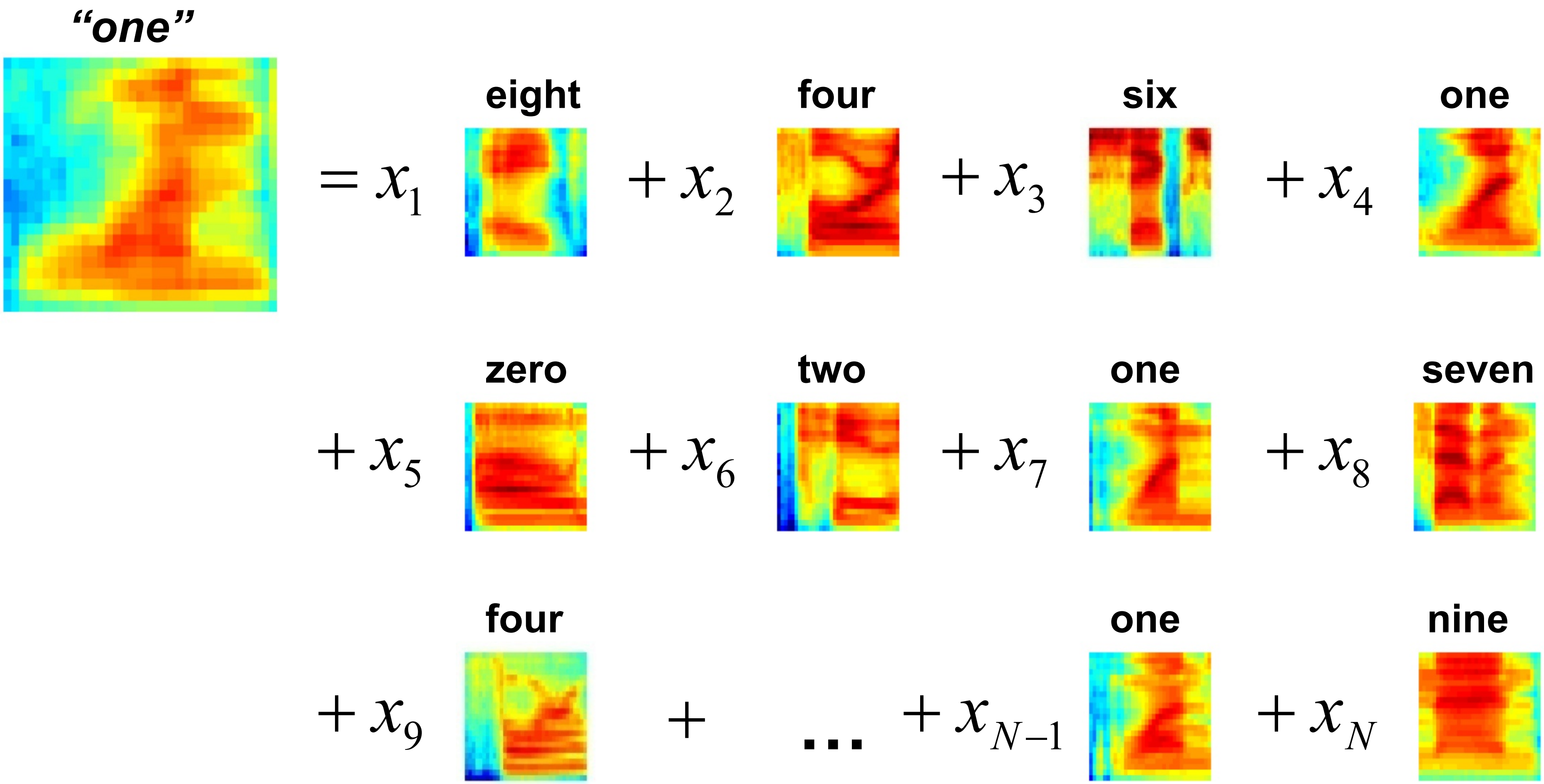
- P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in Proceedings of Neural Information Processing Systems, Vancouver, Canada, 2009.
- J. F. Gemmeke, H. Van hamme, B. Cranen, and L. Boves, “Compressive sensing for missing data imputation in noise robust speech recognition,” IEEE Journal of Selected Topics in Signal Processing, vol. 4, no. 2, pp. 272–287, 2010.

Dictionary sampling

- Approach: directly use samples from the training data
  - Often called “exemplars”
  - This may lead to very large dictionaries!
- Common techniques:
  - random sampling: a random subset of the exemplars
  - pruning: select a subset using some criterion
    - Correlation between exemplars
    - How often an exemplar is activated on development data
- Pros and cons
  - Dictionaries do not always generalize well to unseen data
  - Does not consider additivity: large dictionaries (but activations are *sparse*)
  - Dictionaries are discriminative between sources
  - Simpler to use more time-context (many features)
  - Requires little tuning

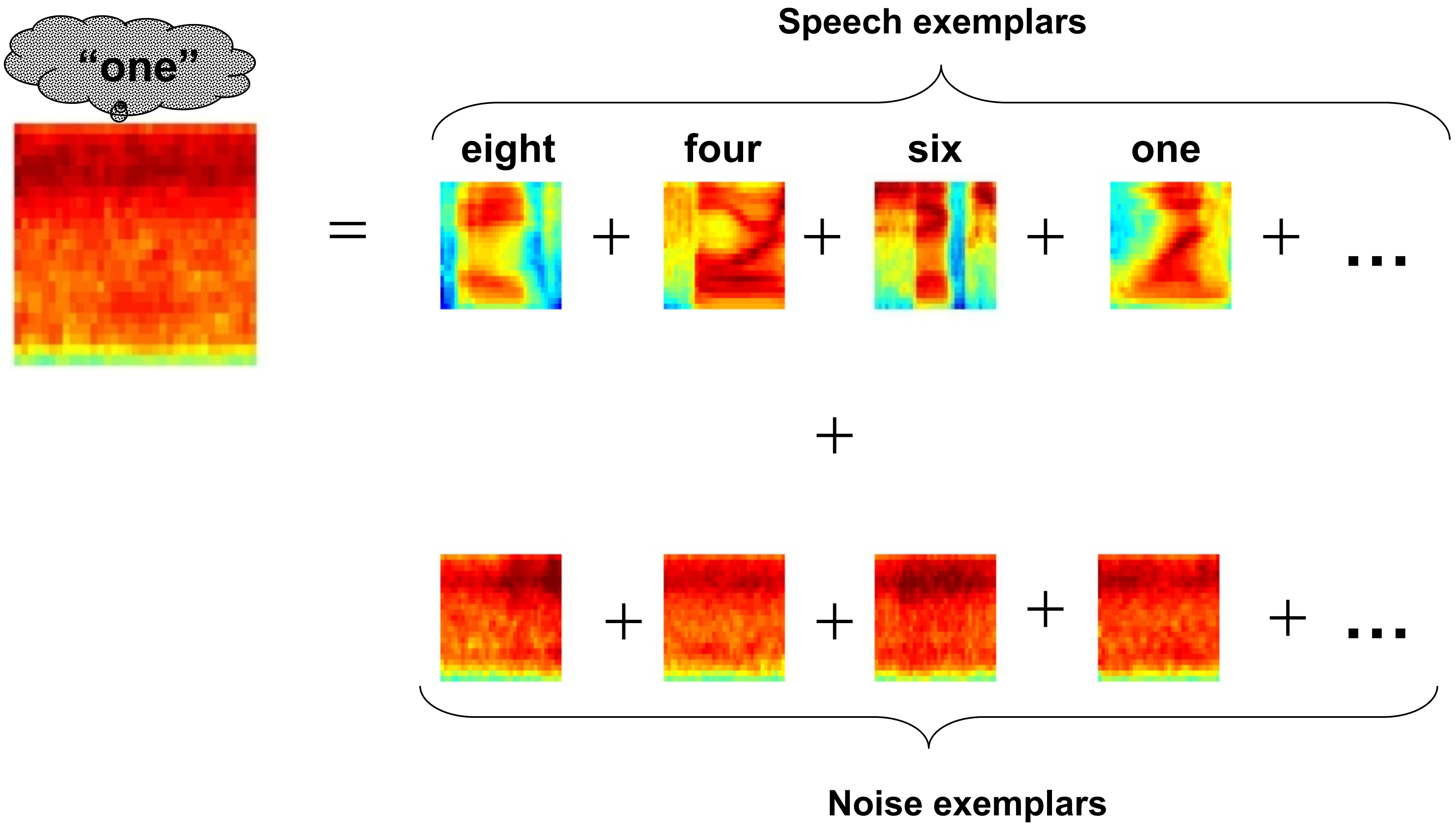
- P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in Proceedings of Neural Information Processing Systems, Vancouver, Canada, 2009.
- J. F. Gemmeke, H. Van hamme, B. Cranen, and L. Boves, “Compressive sensing for missing data imputation in noise robust speech recognition,” IEEE Journal of Selected Topics in Signal Processing, vol. 4, no. 2, pp. 272–287, 2010.

Exemplar-based dictionary



J. F. Gemmeke, H. Van hamme, B. Cranen, and L. Boves, “Compressive sensing for missing data imputation in noise robust speech recognition,” IEEE Journal of Selected Topics in Signal Processing, vol. 4, no. 2, pp. 272–287, 2010.

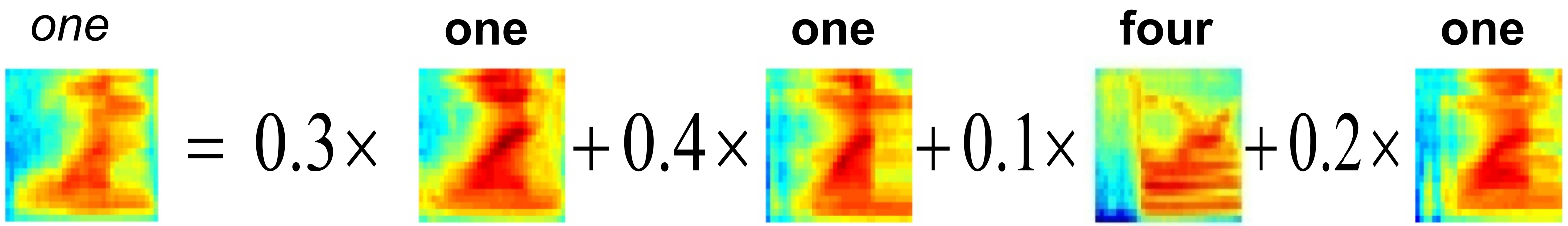
Exemplar-based source separation



J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2067 - 2080, 2011.

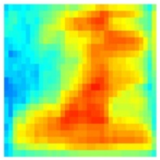
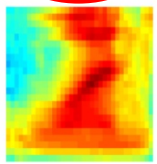
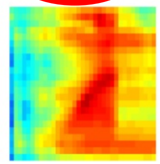
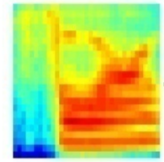
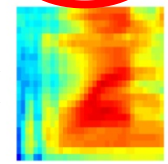
Analyzing the semantics of audio

- Supervised dictionary allows using meta information about each atom

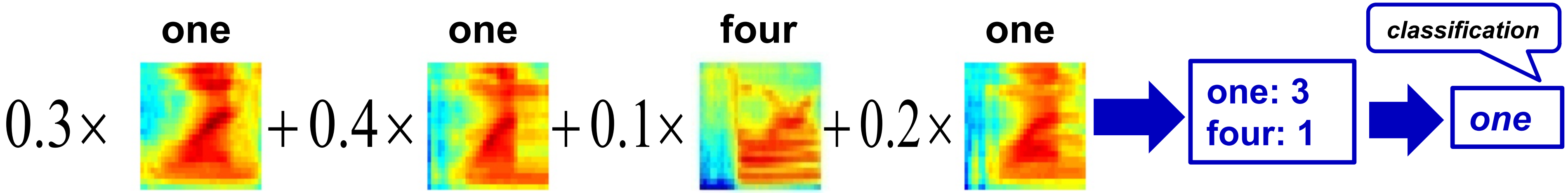


Speech recognition

one **one** **one** **four** **one**

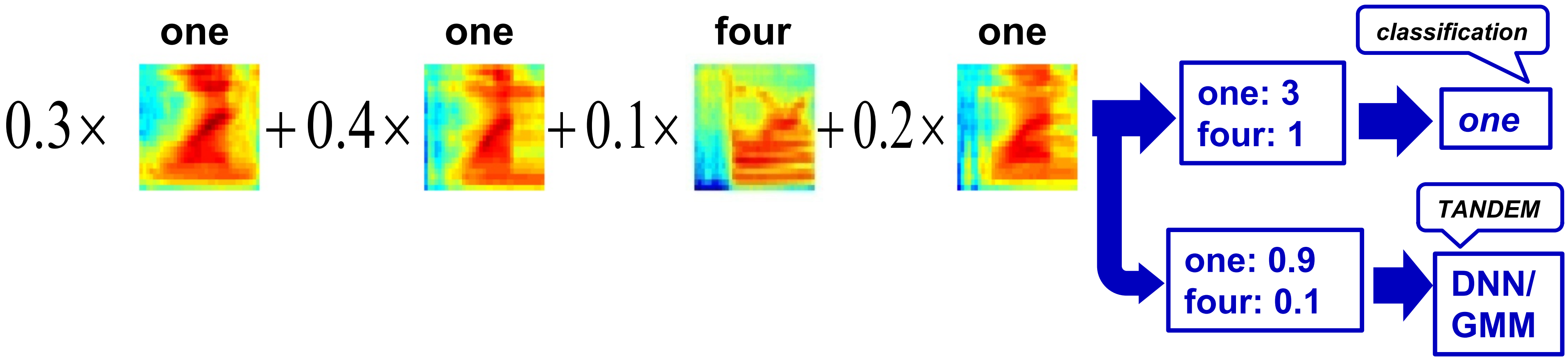
 $= 0.3\times$  $+0.4\times$  $+0.1\times$  $+0.2\times$ 

Speech recognition



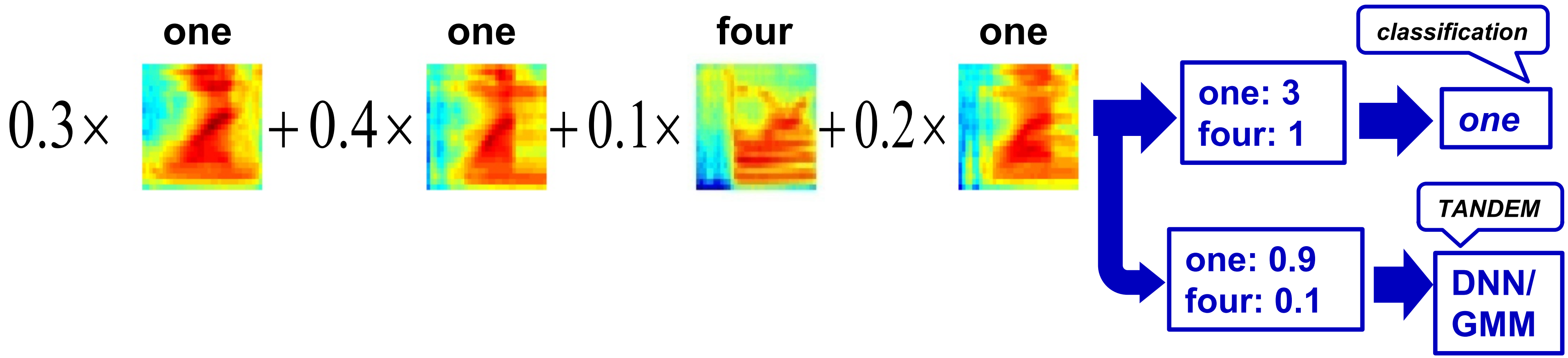
J. F. Gemmeke and L. ten Bosch and L. Boves and B. Cranen, “Using sparse representations for exemplar based continuous digit recognition” In Proc. EUSIPCO, pp. 1755-1759, 2009.

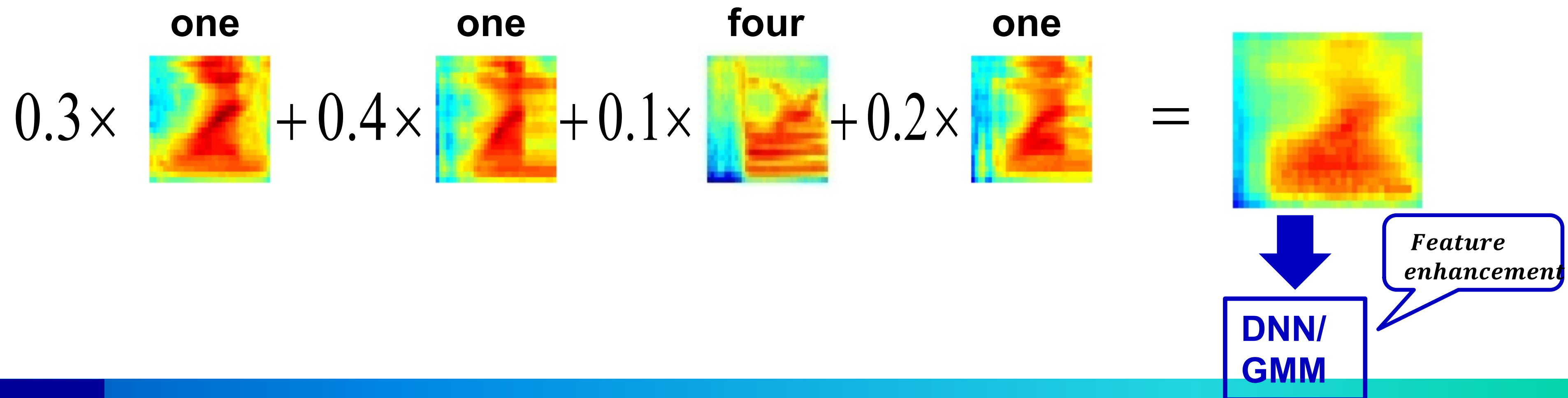
Speech recognition



Y. Sun, B. Cranen, J. F. Gemmeke, L. Boves, L. ten Bosch, and M. M. Doss, "Using sparse classification outputs as feature observations for noise-robust ASR," in Interspeech 2012, Portland, USA, 2012.

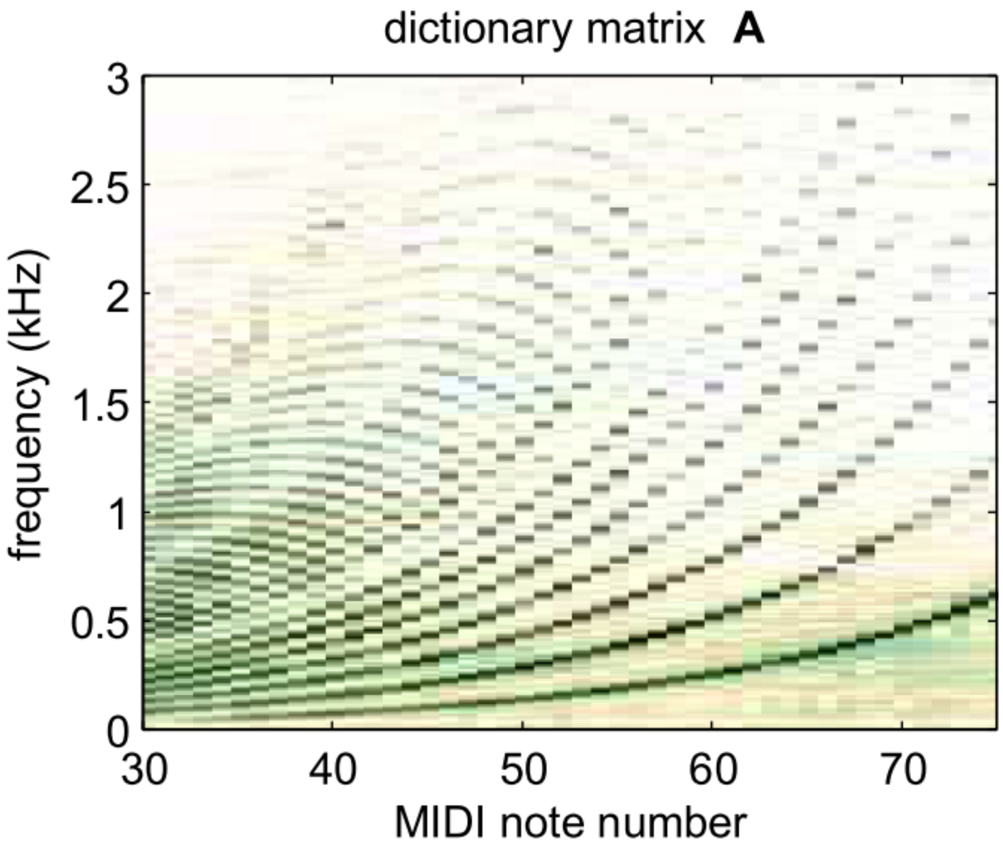
Speech recognition





J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2067 - 2080, 2011.

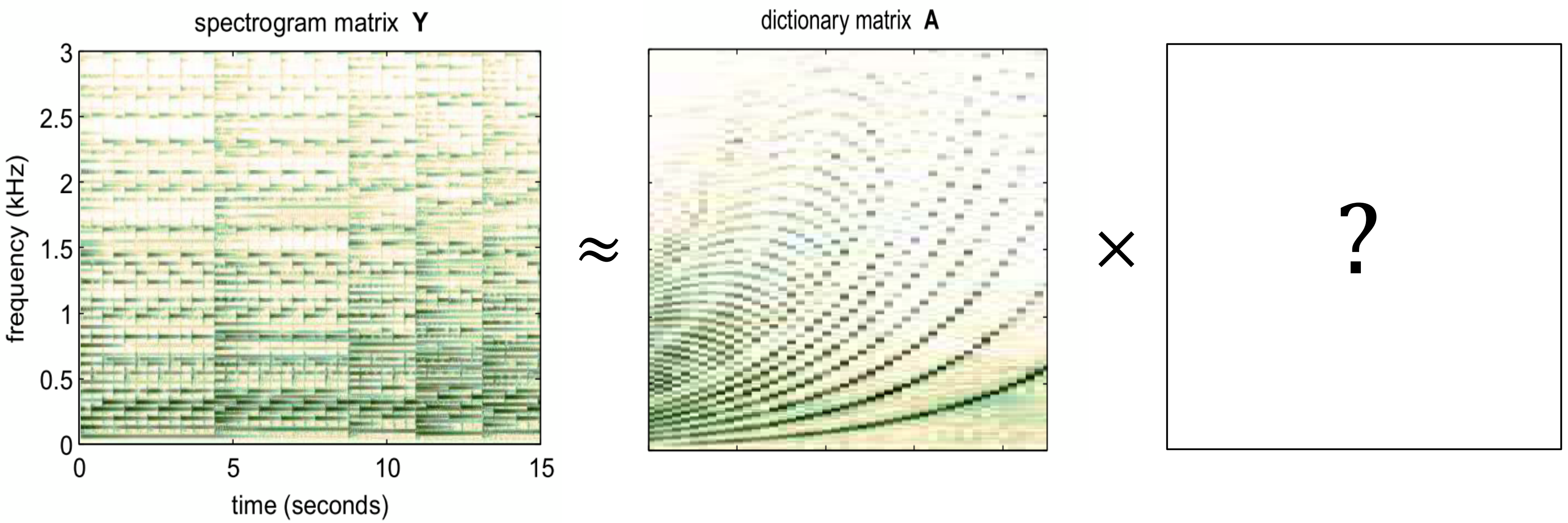
Music example



each atom corresponds to the spectrum of a piano note

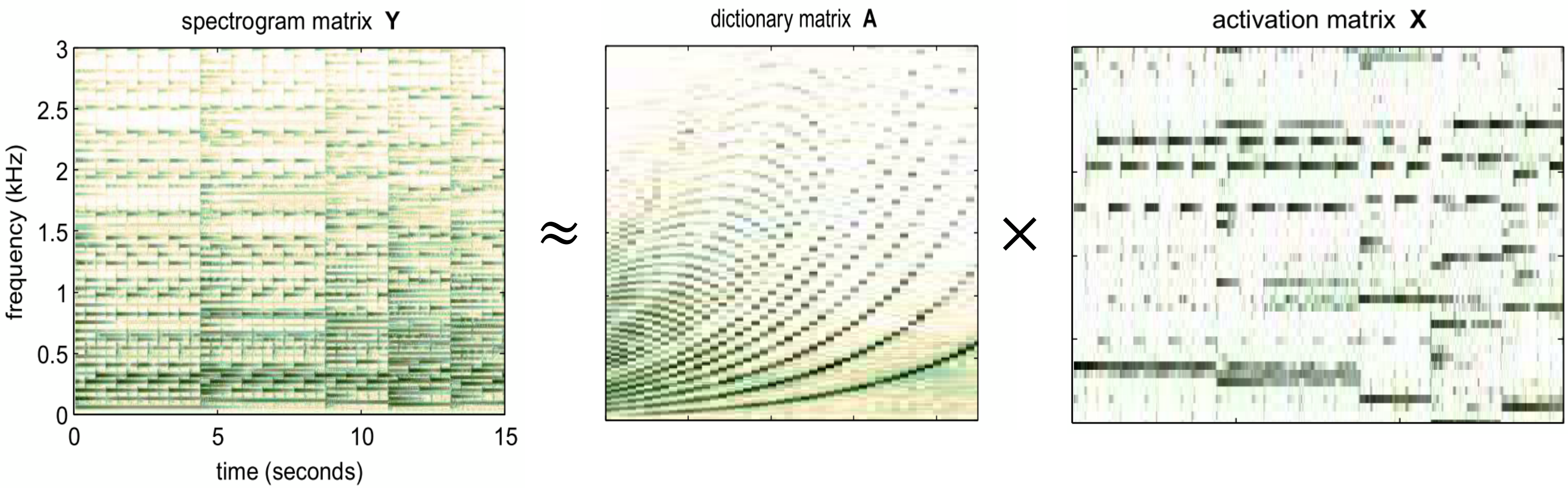
- N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 3, pp. 538 - 549, 2010.
- T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in Proceedings of International Conference on Music Information Retrieval, Kobe, Japan, 2009.

Music analysis problem



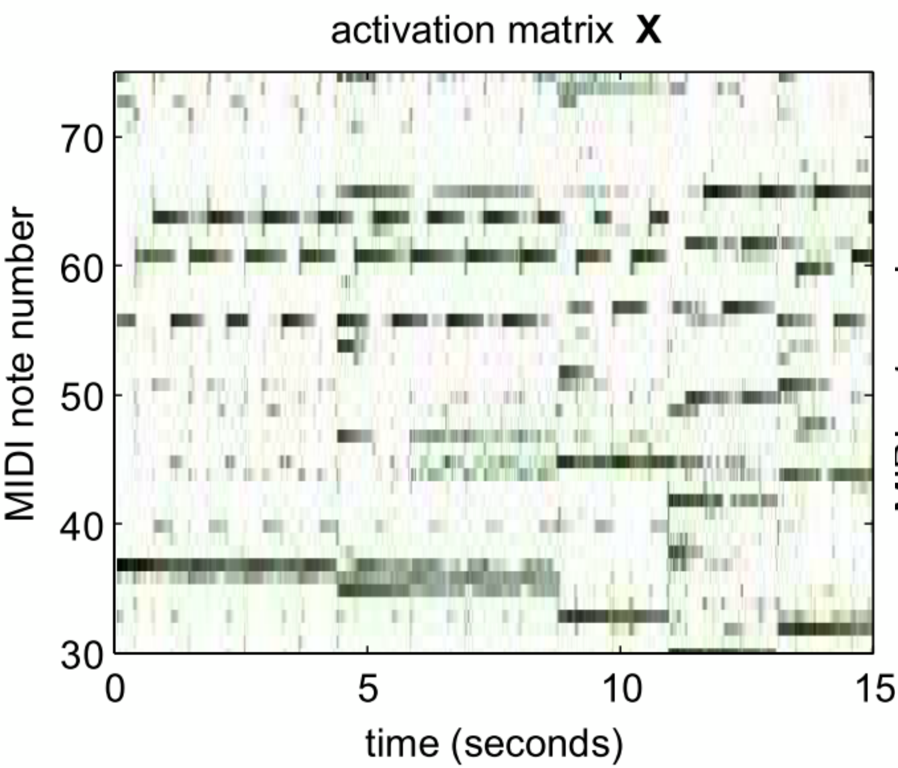
- N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 3, pp. 538 - 549, 2010.
- T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in Proceedings of International Conference on Music Information Retrieval, Kobe, Japan, 2009.

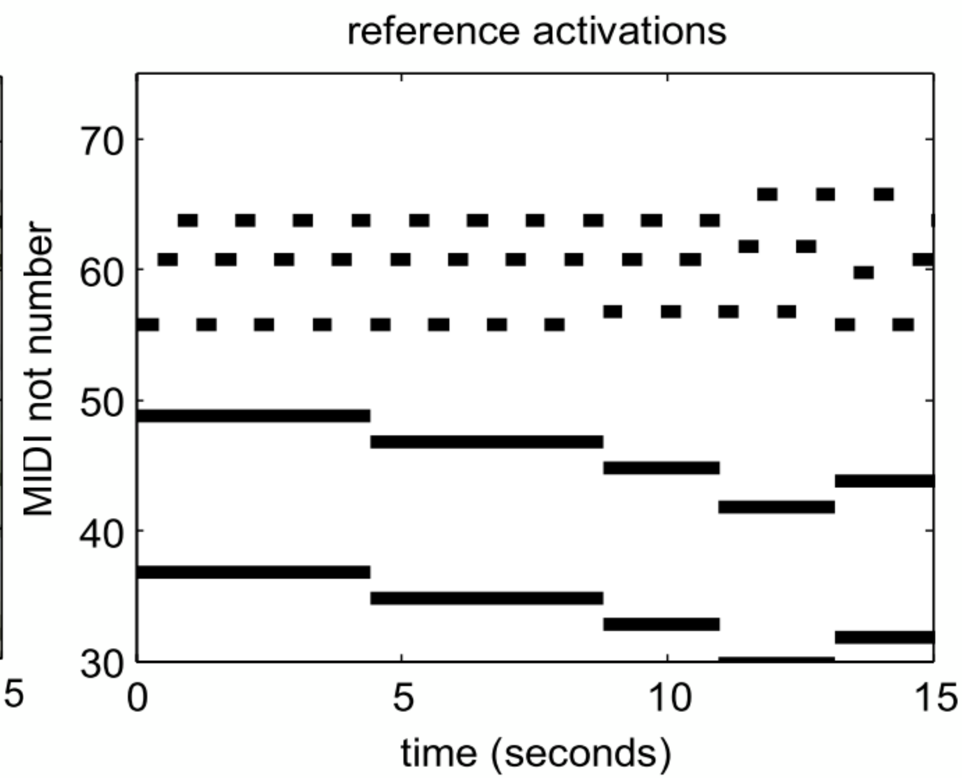
Music analysis problem



- N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 3, pp. 538 - 549, 2010.
- T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in Proceedings of International Conference on Music Information Retrieval, Kobe, Japan, 2009.

Music analysis problem





- N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 3, pp. 538 - 549, 2010.
- T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in Proceedings of International Conference on Music Information Retrieval, Kobe, Japan, 2009.

Regularization in NMF

- Place additional constraints on the NMF formulation:

$$\mathbf{A}^*, \mathbf{X}^* = \underset{\mathbf{A},\mathbf{X}}{\operatorname{argmin}}\, \mathrm{D}(\mathbf{Y}||\mathbf{A}\mathbf{X}) + \lambda \Phi(\mathbf{X})$$

- Leading to modified multiplicative updates:

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\mathbf{A}^\top \frac{\mathbf{Y}}{\mathbf{A}\mathbf{X}}}{\mathbf{A}^\top \mathbf{1} + \lambda \Phi'(\mathbf{X})}$$

- With $\Phi'(\mathbf{X})$ the matrix derivative of $\Phi(\mathbf{X})$ with respect to **X**
- This necessitates an *l*-2 normalization of the columns of **A**

Google: “NMF regularization” ☺

Model alternatives

Google: "NMF regularization" ☺

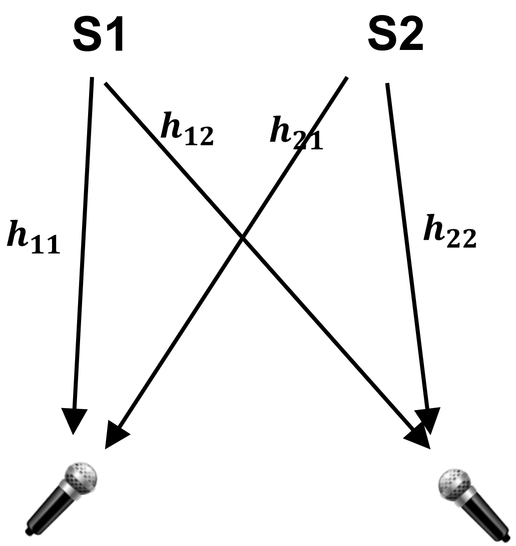
Regularization in NMF

- A very popular regularizer is *sparsity*: $\Phi(\mathbf{X}) = ||\mathbf{X}||_1$
- Sparsity regularisation allows decomposition with overcomplete dictionaries
- Other commonly used regularizers:
  - Temporal continuity (Virtanen 2007)
  - Correlation of weights (Wilson et al. 2008)
  - Correlation of spectra (Virtanen & Cemgil 2009)
  - Correlation of components (Wilson & Raj 2010)
  - Hidden Markov Models (Gemmeke et. al. 2013)

- J. Eggert and E. Korner, “Sparse coding and NMF,” in IEEE International Joint Conference on Neural Networks, pp. 2529–2533, 2004
- P. O. Hoyer, “Non-negative matrix factorization with sparseness constraints,” Journal of Machine Learning Research, vol. 5, pp. 1457–1469, 2004.
- 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, pp. 1066 – 1074, 2007.

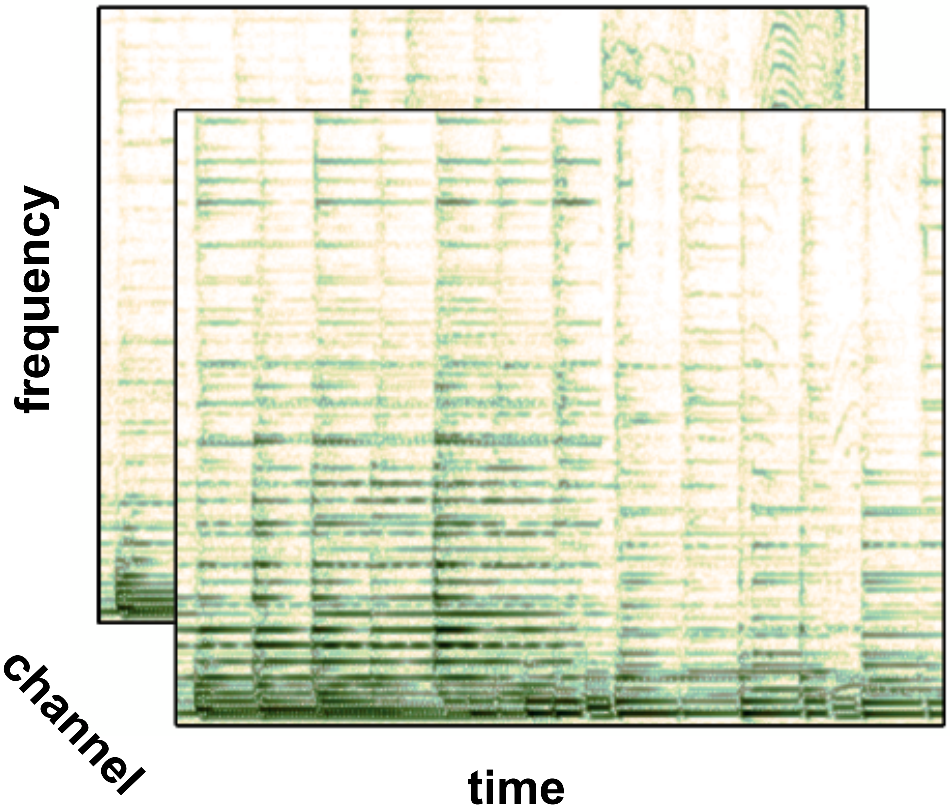
Multichannel audio

- How to use information about differences between channels?
- Phase differences: cannot be modeled with compositional models, require additional modeling components
- Possible to model amplitude differences using compositional models



Multichannel audio

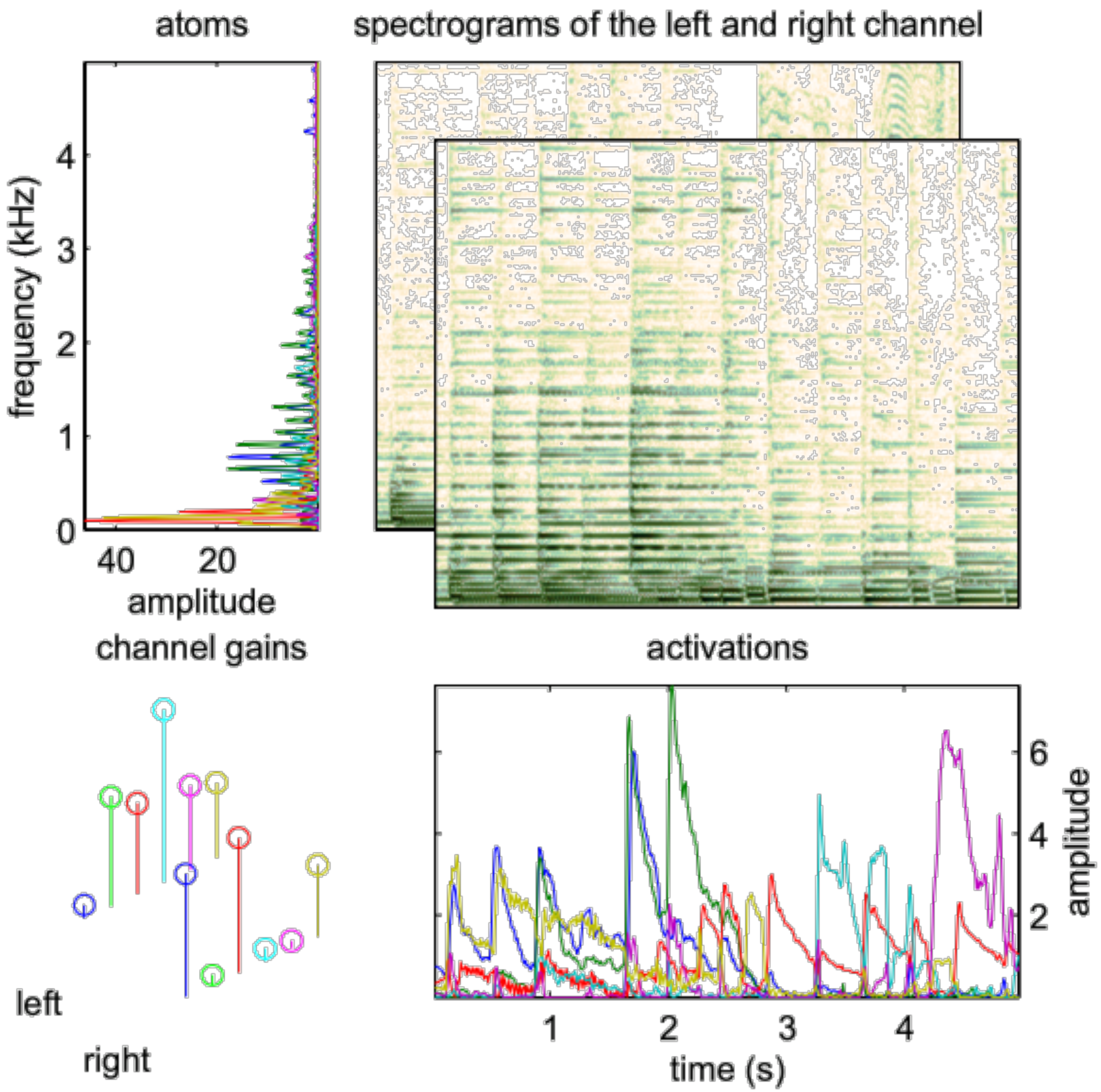
- Signal of each channel represented using spectrogram
- Combined into a 3-D *tensor*



Multichannel audio

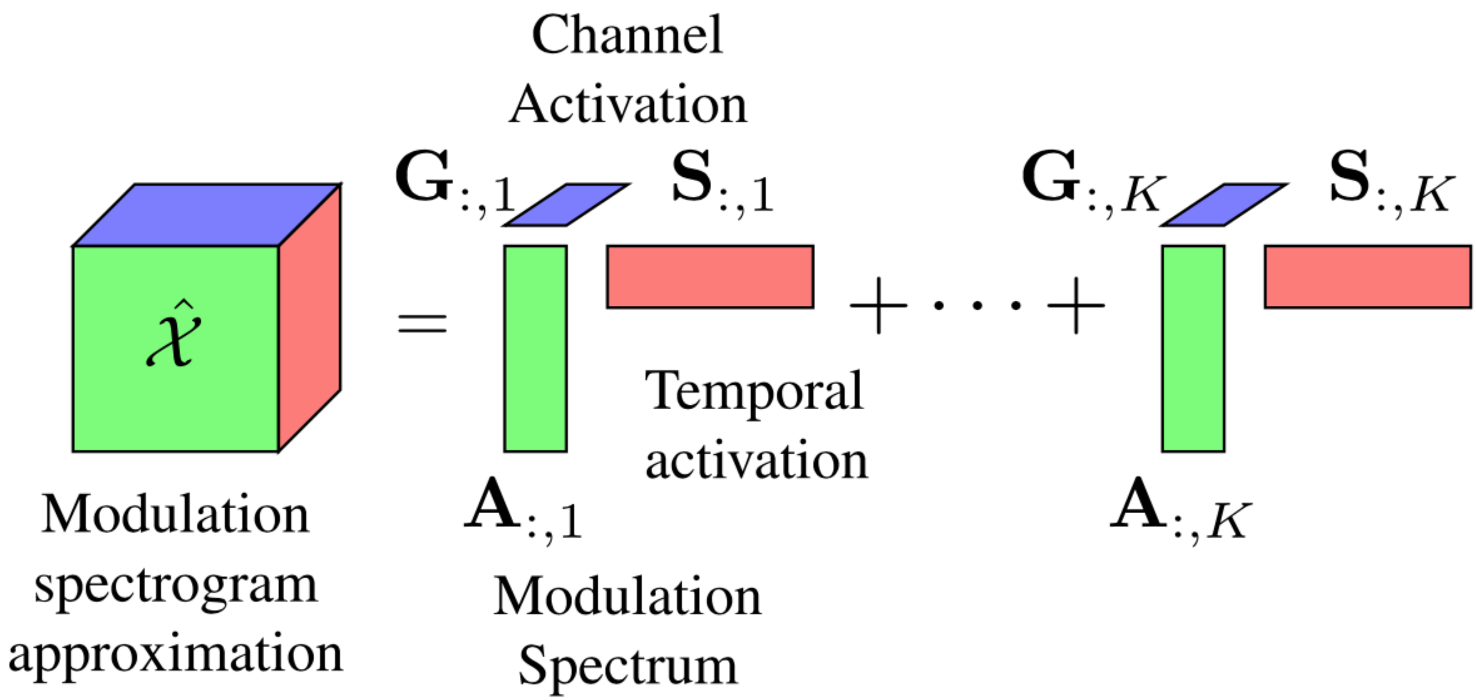
- Tensor decomposed into a sum of components
- Each component presented as an outer product of

1. Spectral atoms
2. Temporal activations
3. Channel gains



- A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," IEEE Trans. on Audio, Speech and Lang. Proc., vol. 18, no. 3, pp. 550-563, March 2010.
- H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Formulations and algorithms for multichannel complex NMF," in Proceedings of IEEE International Conference on Audio, Speech and Signal Processing, Prague, Czech Republic, 2011.

Tensor factorisation of modulation spectrograms



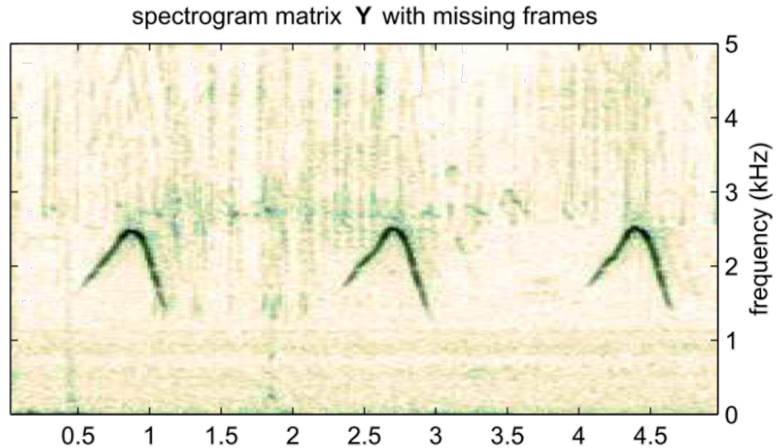
- T. Barker, T. Virtanen, "Blind separation of audio mixtures through nonnegative tesnsor factorisation of modulation spectrograms", in IEEE/ACM Transactions on Audio, Speech and Language Processing, Volume 24, Issue 12, December 2016, pp. 2377-2389.

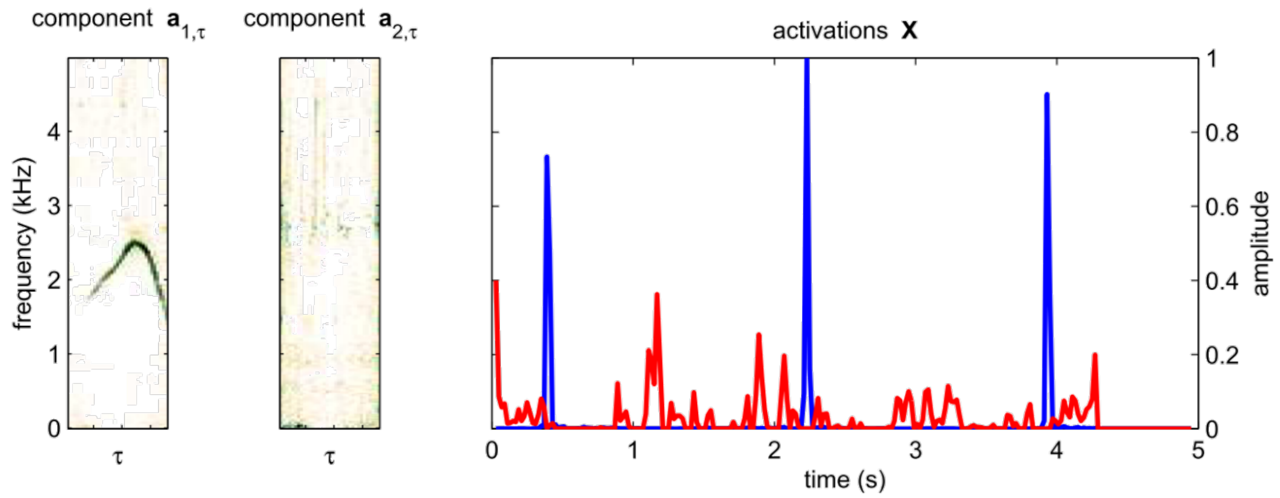
Temporal context

- Sound typically has much spectral and temporal structure
- The basic NMF model treats each frequency and frame as independent from each other
- Modelling contextual information (time-frequency patches) useful

Temporal context: NMF deconvolution

$$\hat{\mathbf{y}}_t = \sum_k \sum_\tau \mathbf{a}_{k,\tau} x_{k,t-\tau}$$



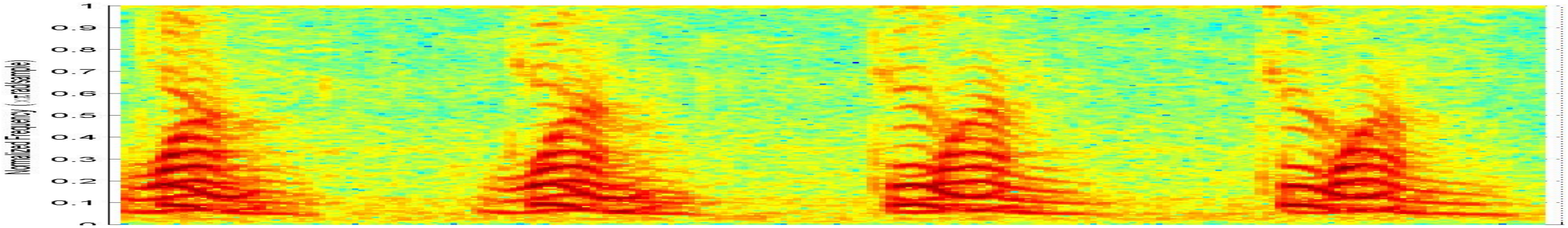


- P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 1, pp. 1 – 12, 2007.
- P. D. O. Grady, "Sparse separation of underdetermined speech mixtures," Ph.D. dissertation, National University of Ireland, Maynooth, 2007.

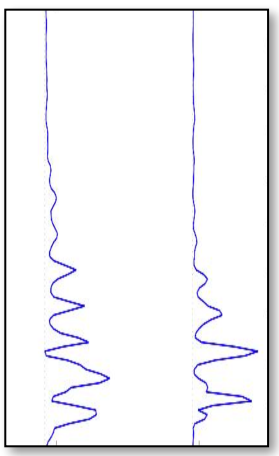
Example

- Two distinct sounds occurring with different repetition rates within a signal

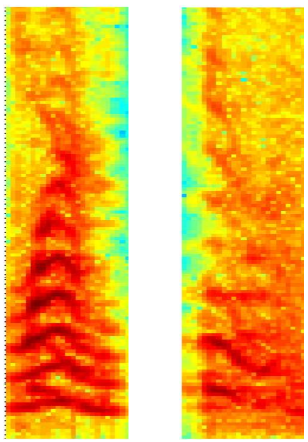
INPUT SPECTROGRAM



NMF dictionary atoms

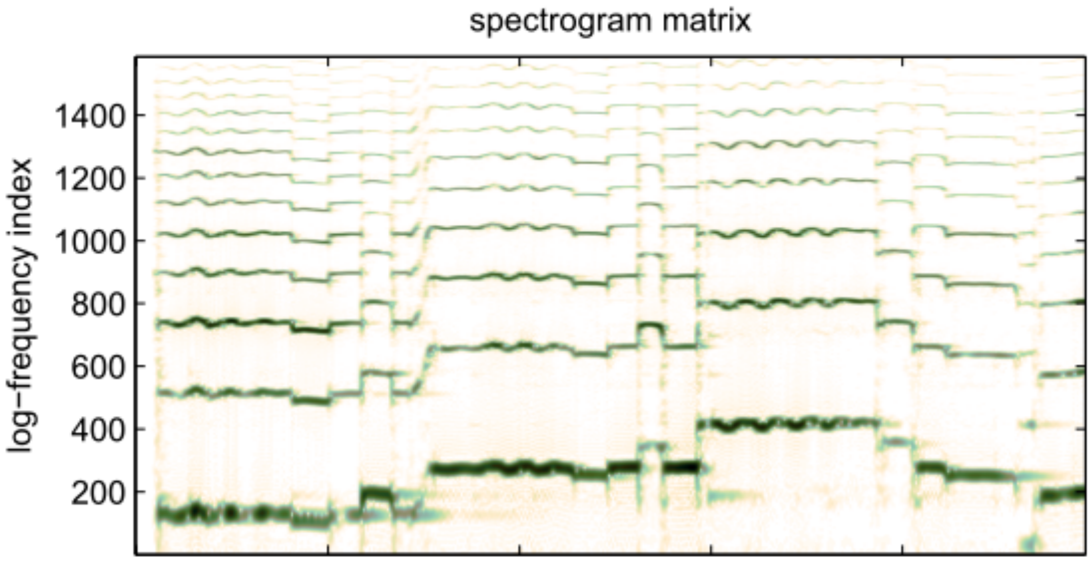


Discovered time-frequency "patch" atoms



Invariant models

- The basic NMF model requires a separate atom to model sounds with different pitch
- Modeling multiple pitches with a single atom?
- On a log-frequency scale, pitch shifting corresponds to translation

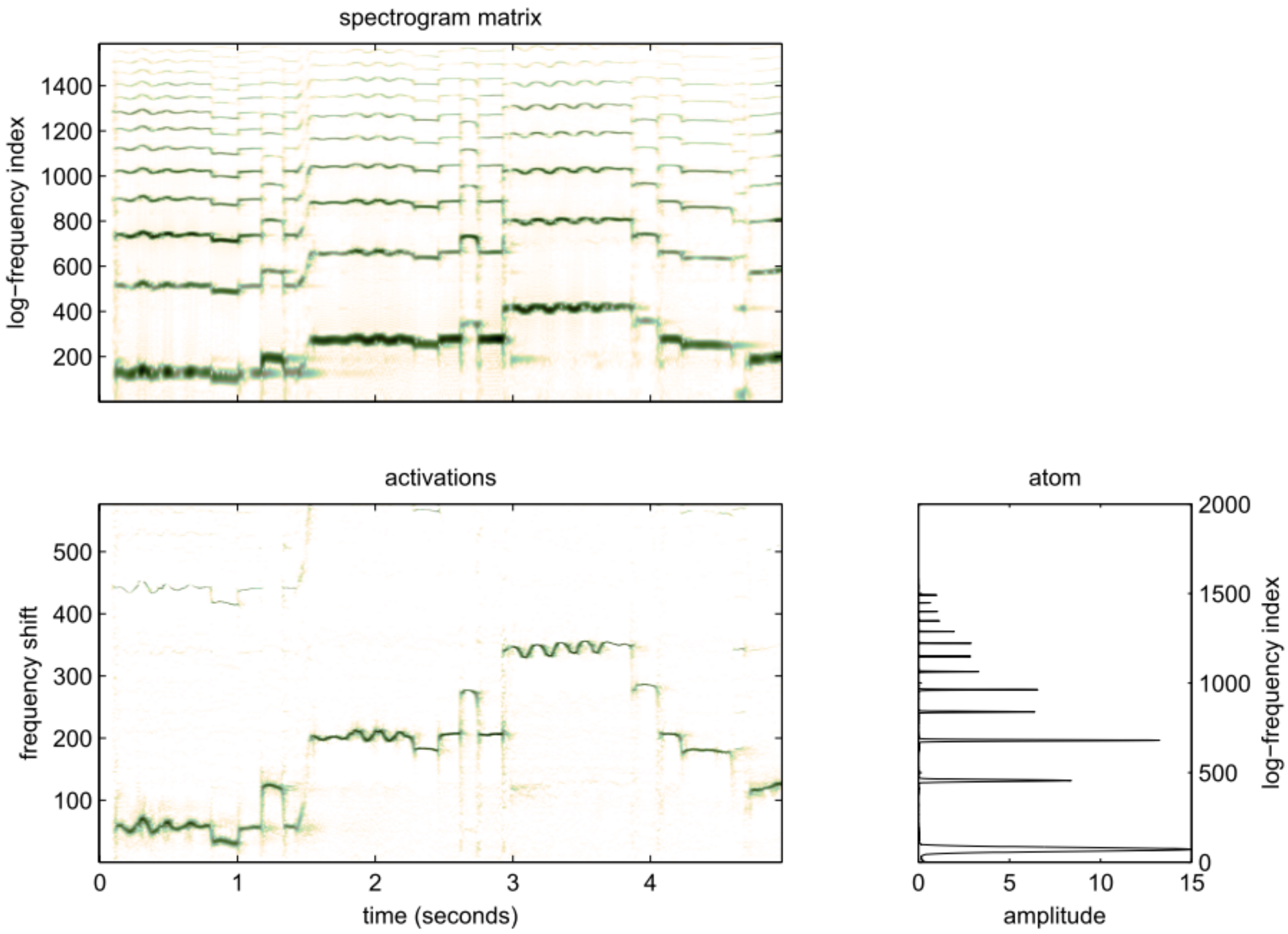


NMF deconvolution in frequency

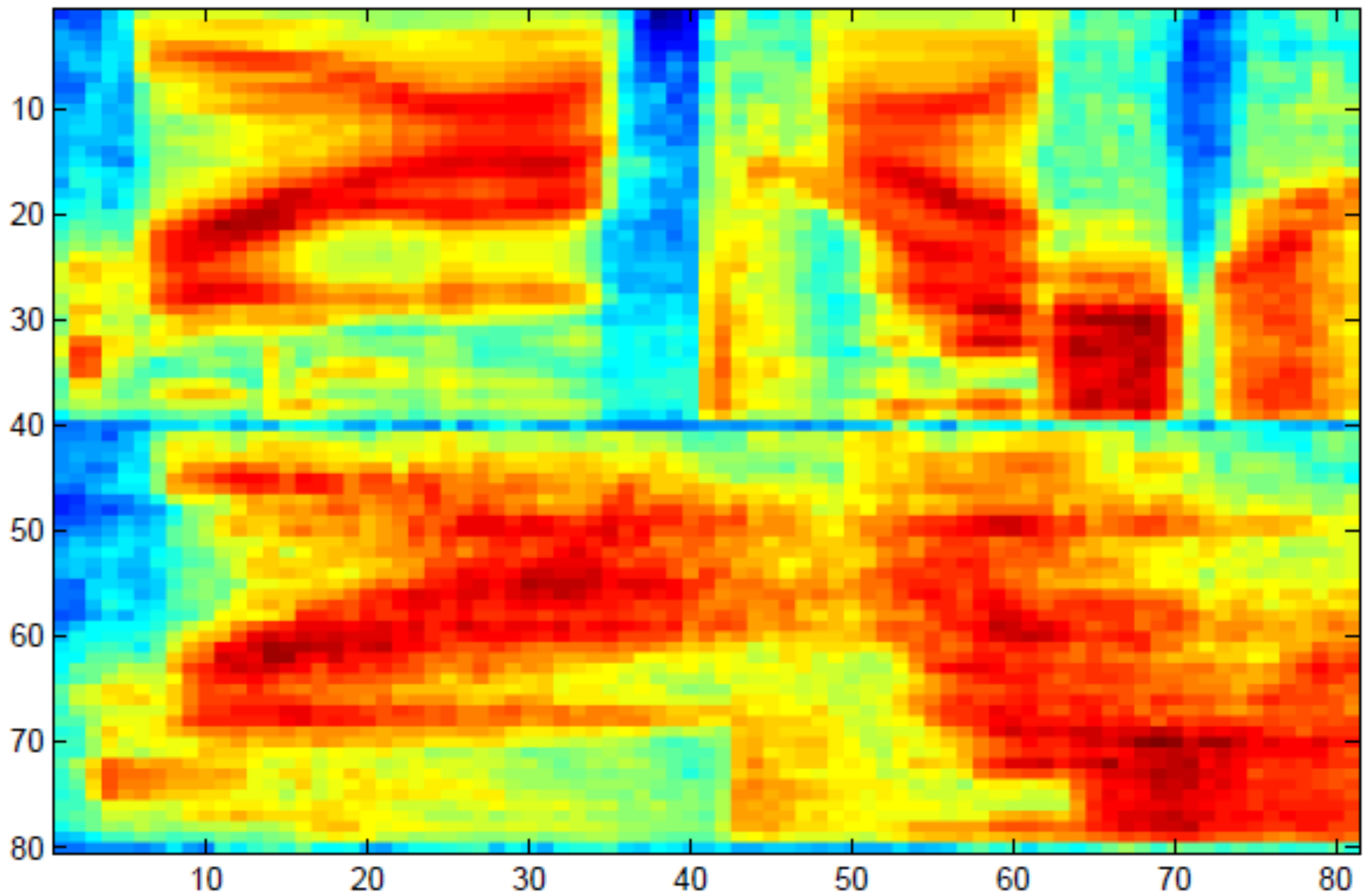
- Basic idea:
  - Use logarithmic frequency axis
  - Allow translating atoms a in frequency (convolution)
  - Estimate activation x_τ for each amount τ of translation

$$\hat{y}_{f,t} = \sum_{k=1}^{K} \sum_{\tau \in \mathcal{L}} a_{f+\tau,k} x_{k,\tau}$$

NMF deconvolution

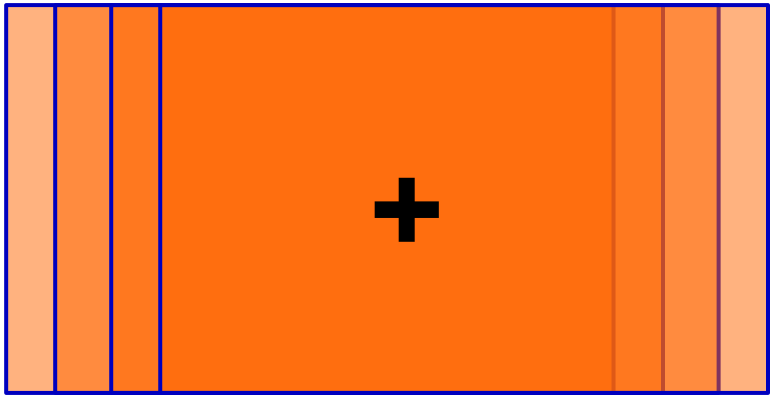


Dereverberation

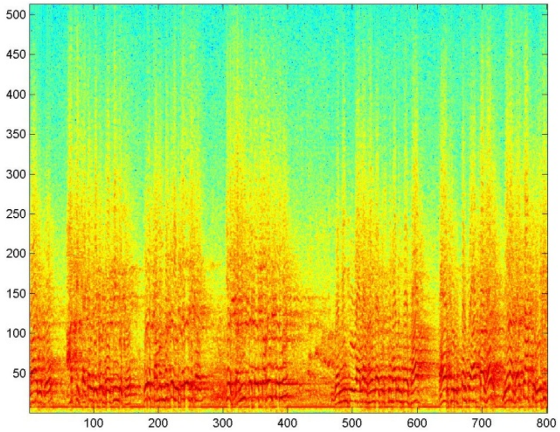


Dereverberation

- A convolutional model of reverberation:
  - The spectrogram of the reverberated signal is a sum of the spectrogram of the clean signal and several shifted and scaled versions of itself

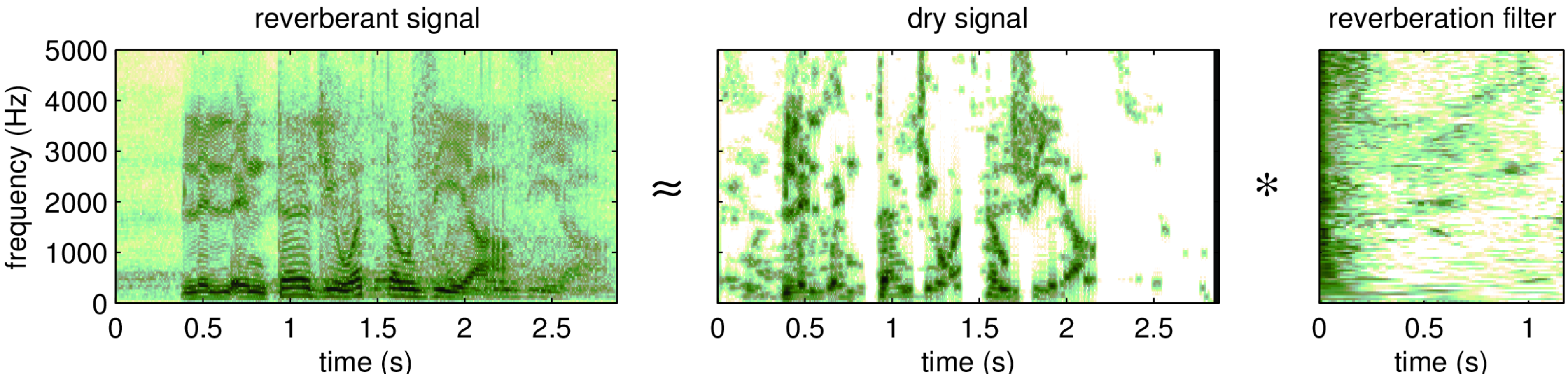


=



Dereverberation

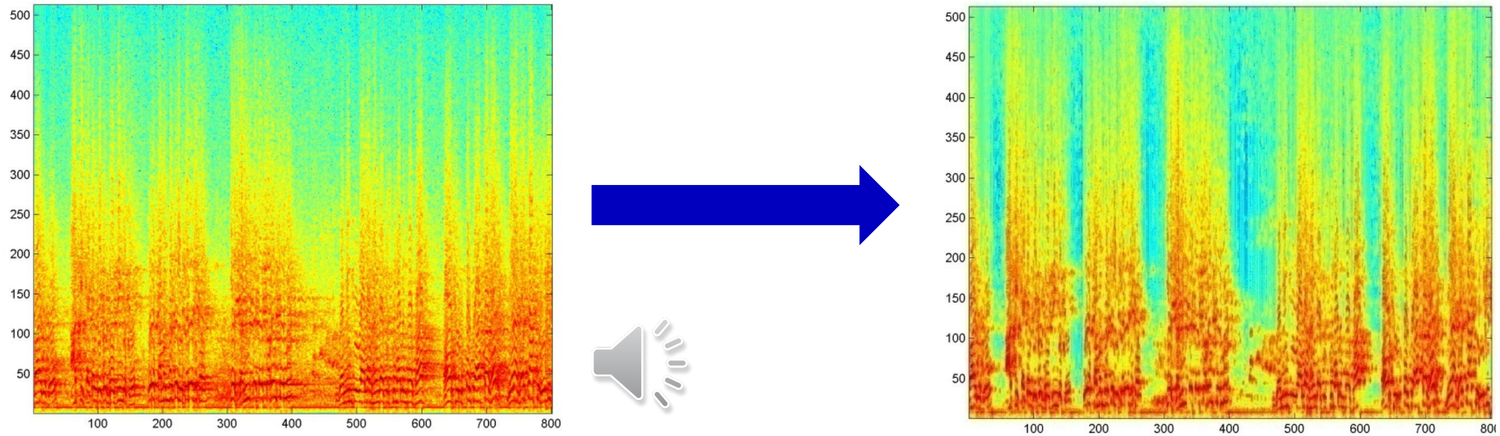
- A convolutional model of reverberation:
  - The spectrogram of the reverberated signal is a sum of the spectrogram of the clean signal and several shifted and scaled versions of itself
  - A convolution of the spectrogram and a room response



- N. Yasuraoka, H. Kameoka, T. Yoshioka, and H. G. Okuno, "I-divergence-based dereverberation method with auxiliary function approach," in Proceedings of IEEE International Conference on Audio, Speech and Signal Processing, Prague, Czech Republic, 2011.
- R. Singh, B. Raj, and P. Smaragdis, "Latent-variable decomposition based dereverberation of monaural and multi-channel signals," in Proceedings of IEEE International Conference on Audio, Speech and Signal Processing, Dallas, USA, 2010.

Dereverberation

- A convolutional model of reverberation:
  - The spectrogram of the reverberated signal is a sum of the spectrogram of the clean signal and several shifted and scaled versions of itself
  - A convolution of the spectrogram and a room response
  - Factorial model: **Y=SH**, with **Y** the reverberated spectrum, **S** the dry spectrum, and **H** the reverberation filter
    - Sparsity must be enforced on the filter



- N. Yasuraoka, H. Kameoka, T. Yoshioka, and H. G. Okuno, "I-divergence-based dereverberation method with auxiliary function approach," in Proceedings of IEEE International Conference on Audio, Speech and Signal Processing, Prague, Czech Republic, 2011.
- R. Singh, B. Raj, and P. Smaragdis, "Latent-variable decomposition based dereverberation of monaural and multi-channel signals," in Proceedings of IEEE International Conference on Audio, Speech and Signal Processing, Dallas, USA, 2010.

Comparison to DNNs

- DNNs are discriminative models
  - Ideal for classification
- Compositional models are generative
  - Can explain the properties of the data better

Comparison to DNNs

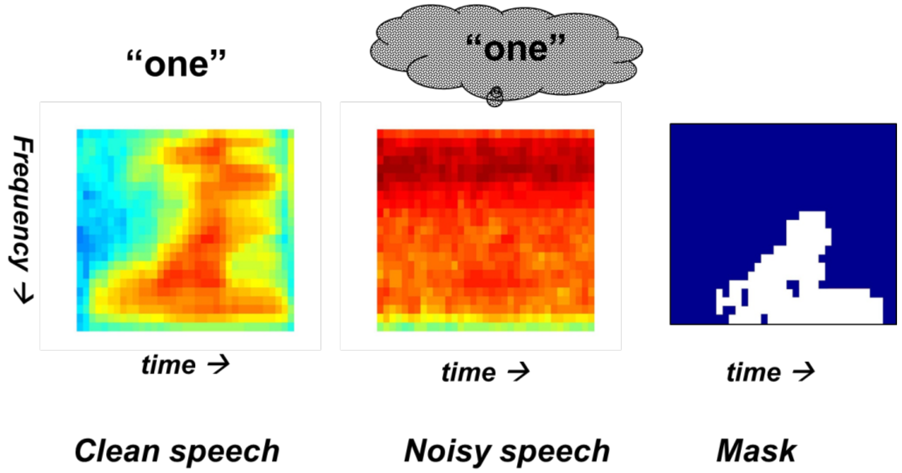
- DNNs give state of the art results in many application areas
  - Applicable also on many mentioned problems (source separation, robust recognition)
- Given large amounts of training data, DNNs typically outperform compositional models
- Compositional models can be used with small amounts of training data
  - Examplar-based dictionaries obtained from few examples
- Compositional models enable unsupervised processing that does not require training data

Missing data

- Missing data occurs in many applications
  - Packet or frame drops
  - Signal clipping
  - Audio corrupted at specific frequencies

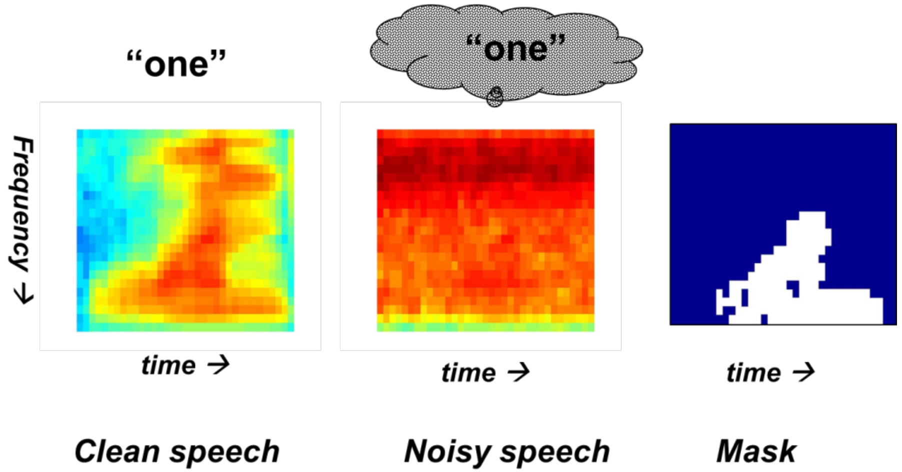
Missing data

- Missing data occurs in many applications
  - Packet or frame drops
  - Signal clipping
  - Audio corrupted at specified frequencies



Missing data

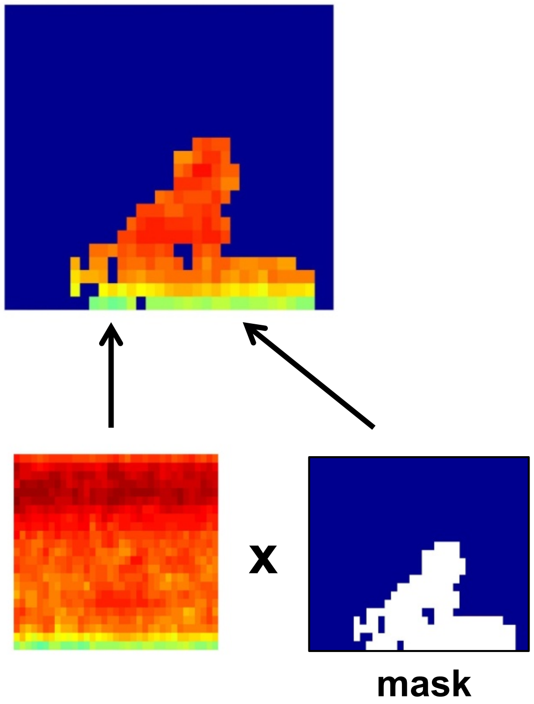
- Missing data occurs in many applications
  - Packet or frame drops
  - Signal clipping
  - Audio corrupted at specified frequencies



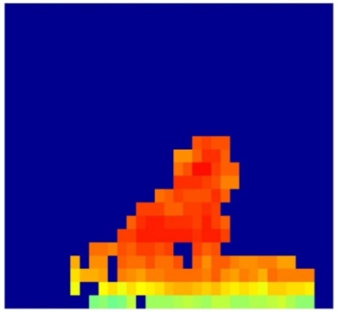
- Original audio: $\mathbf{s} \approx \hat{\mathbf{s}} = \mathbf{A}\mathbf{x}$
- Missing data: $\mathbf{M}\mathbf{y} \approx \mathbf{M}\hat{\mathbf{s}} = \mathbf{M}\mathbf{A}\mathbf{x}$

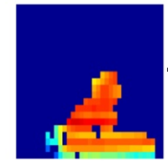
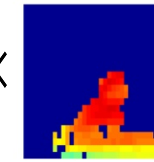
- Smaragdis, P., B. Raj, M. Shashanka, Missing data imputation for spectral audio signals, in IEEE international workshop on Machine Learning for Signal Processing (MLSP), 2009
- J. F. Gemmeke, H. Van hamme, B. Cranen, and L. Boves, "Compressive sensing for missing data imputation in noise robust speech recognition," IEEE Journal of Selected Topics in Signal Processing, vol. 4, no. 2, pp. 272–287, 2010.

Clean speech reconstruction

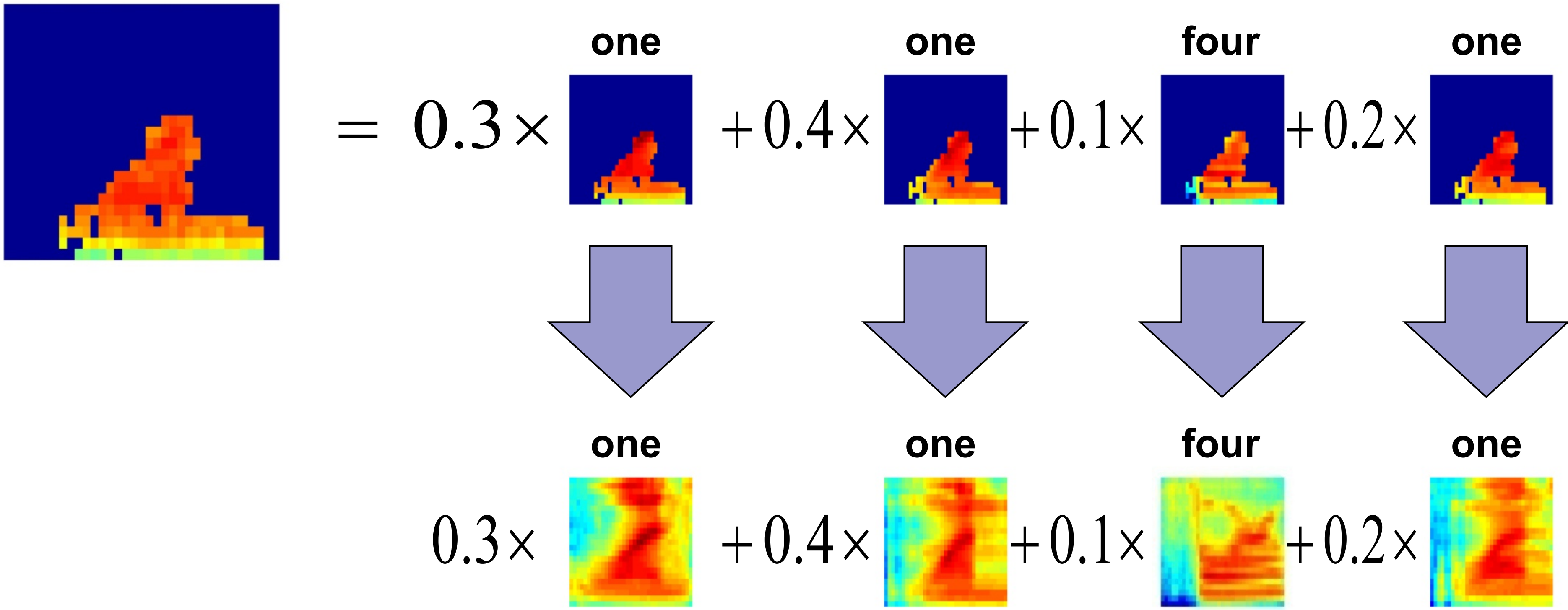


Clean speech reconstruction

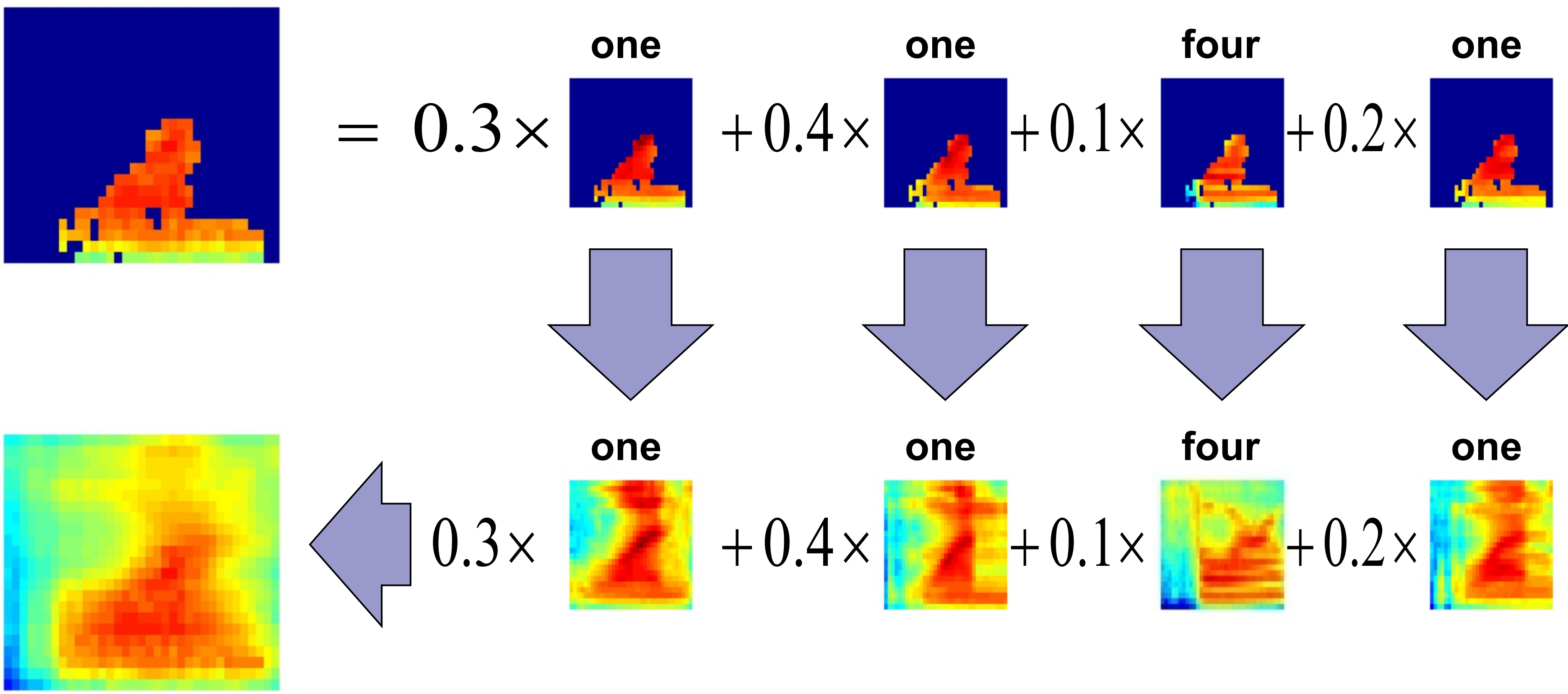


$=0.3\times$ one  $+0.4\times$ one  $+0.1\times$ four  $+0.2\times$ one 

Clean speech reconstruction

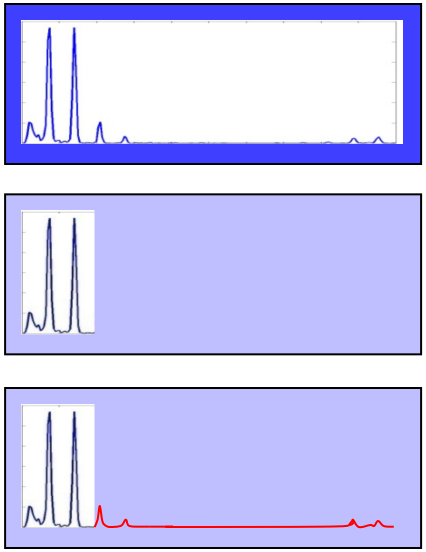


Clean speech reconstruction



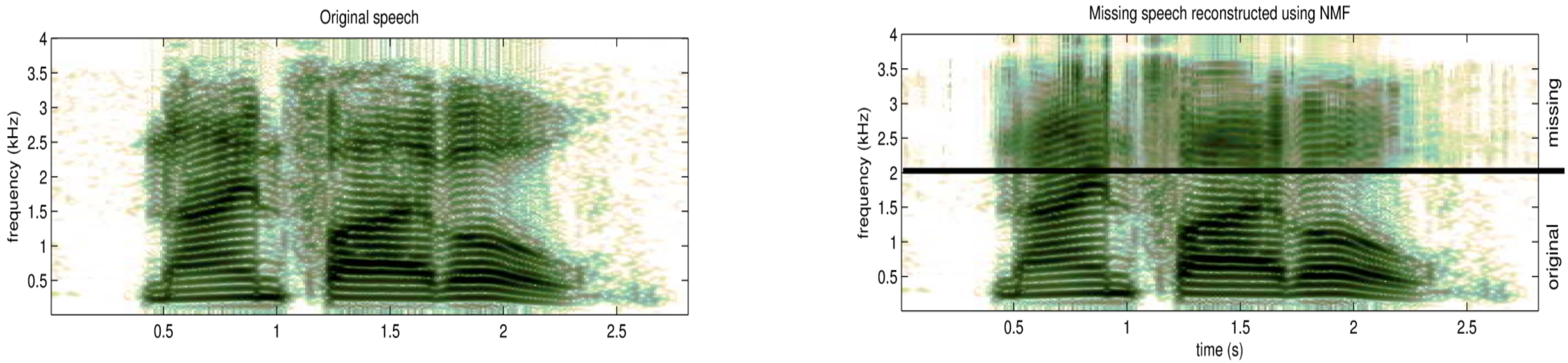
Bandwidth extension

- Problem: A given speech signal only has frequencies in the 300Hz-3.5Khz range
  - Telephone quality speech 
- Goal: restore the missing frequencies
- Assumptions:
  - We have full-bandwidth training data
  - Training data is representative
  - We know which frequencies are missing



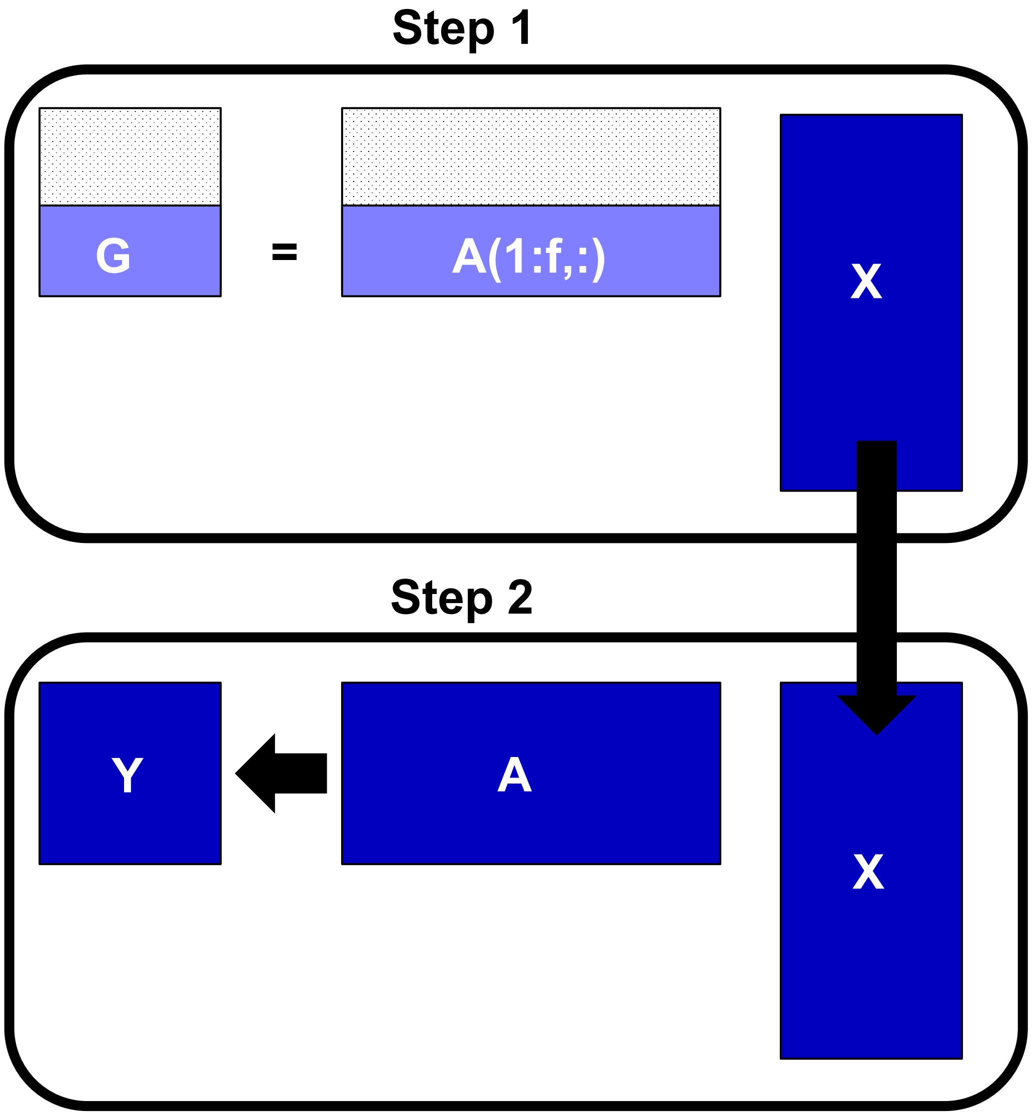
Bandwidth extension

- Almost trivial use of compositional models (using Matlab notation):
  - Step 1: Create a full-bandwidth dictionary **A**
  - Step 2: Given the limited bandwidth observation **G,** in which only the frequency bands 1…f are retained, use a bandwidth-limited **A**(1:f,:) dictionary to obtain activations **X**
  - Step 3: Reconstruct the full-bandwidth estimate **Y** using the full bandwidth dictionary **A**



D. Bansal, B. Raj, and P. Smaragdis, "Bandwidth expansion of narrowband speech using non-negative matrix factorization," in 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, 2003.

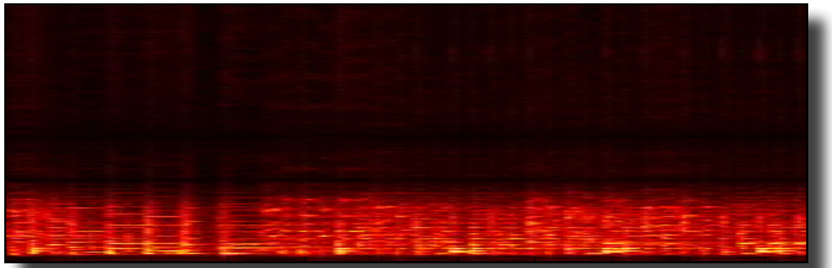
Visual representation



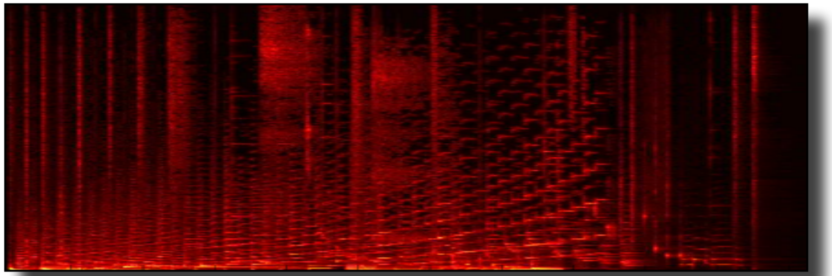
D. Bansal, B. Raj, and P. Smaragdis, "Bandwidth expansion of narrowband speech using non-negative matrix factorization," in 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, 2003.

Audio example

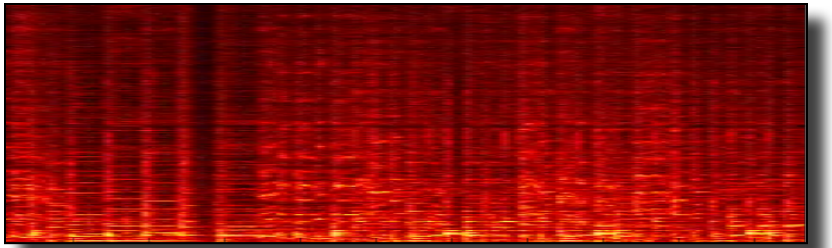
Reduced BW data



Training material



Bandwidth expanded version



Getting started

- Literature
  - Check the references ☺
  - Tutorial article: T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis. Compositional Models for Audio Processing. IEEE Signal Processing Magazine, March 2015
- Matlab code
  - Supervised NMF-based SS: http://www.cs.tut.fi/~tuomasv/software.html
  - SS, recognition and imputation: http://www.amadana.nl/software
  - FASST, evalation, etc: http://www.loria.fr/~evincent/soft.html
  - NMFlab: http://www.bsp.brain.riken.jp/ICALAB/nmflab.html
  - PLCA: http://www.cs.illinois.edu/~paris/pubs/
- C++ code
  - openBliSSART: http://openblissart.github.io/openBliSSART/

Summary

- Realistic sounds consist of components that combine purely additively
- Composition models are purely additive models that are powerful in modeling sound mixtures
- Good models for spectral representations of sound
- Applications in source separation and audio processing

Acknowledgements

- I am indebted to *Jort Gemmeke*, *Bhiksha Raj, Paris Smaragdis* and *Meng Sun* for providing much of the source material for these slides

- I thank my collaborators at: KU Leuven, Radboud University Nijmegen, Technical University, Tampere, CMU, Technical University München, Aalto University, IBM, …

The end...