SPECTRAL COVARIANCE IN PRIOR DISTRIBUTIONS OF NON-NEGATIVE MATRIX FACTORIZATION BASED SPEECH SEPARATION

Tuomas Virtanen

Department of Signal Processing, Tampere University of Technology
Korkeakoulunkatu 1, FI-33720 Tampere, Finland
phone: + (358) 331154798, fax: + (358) 331153857, email: tuomas.virtanen@tut.fi
web: `www.cs.tut.fi/~tuomasv`

ABSTRACT

This paper proposes an algorithm for modeling the covariance of the spectrum in the prior distributions of non-negative matrix factorization (NMF) based sound source separation. Supervised NMF estimates a set of spectrum basis vectors for each source, and then represents a mixture signal using them. When the exact characteristics of the sources are not known in advance, it is advantageous to train prior distributions of spectra instead of fixed spectra. Since the frequency bands in natural sound sources are strongly correlated, we model the distributions with full-covariance Gaussian distributions. Algorithms for training and applying the distributions are presented. The proposed methods produce better separation quality that the reference methods. Demonstration signals are available at `www.cs.tut.fi/~tuomasv`.

1. INTRODUCTION

Sound source separation has applications in the analysis and manipulation of audio signals, since individual sources can be recognized or modified more efficiently than polyphonic mixtures. Recently, non-negative matrix factorization (NMF) has been successfully used in audio source separation [7, 9, 10].

In NMF, the spectrum vector $\mathbf{s}_t^n$ of a sound source n in frame t is approximated as a weighted sum

$$\mathbf{s}_t^n \approx \sum_{i \in \mathcal{S}_n} \mathbf{b}_i g_{t,i} \tag{1}$$

of spectral basis vectors $\mathbf{b}_i$. Here $\mathcal{S}_n$ denotes the set of basis vector indices of source n and $g_{t,i}$ the gain of the ith basis vector in time frame $t, = 1, \ldots, T$. The index sets are disjoint, i.e., each source is presented with a separate set of basis vectors. Both the gains and basis vectors are restricted to be entry-wise non-negative.

The spectral vector $\mathbf{x}_t$ of the mixture signal is a sum of N sources, i.e.,

$$\mathbf{x}_t = \sum_{n=1}^{N} \mathbf{s}_t^n, \tag{2}$$

which results to approximation

$$\mathbf{x}_t \approx \sum_{n=1}^{N} \sum_{i \in \mathcal{S}_n} \mathbf{b}_i g_{t,i}. \tag{3}$$

In the supervised NMF, the basis vectors of each source are trained using trainig data where the source is present in isolation. The basis vectors are then fixed and gains are estimated by minimizing the error of the approximation (3). This procedure can lead to good separation results when appropriate training data is available [7].

When the exact characteristics of a source are not known a priori, it is advantageous to adapt the basis vectors [9]. Instead of fixed basis vectors, we can train prior distributions $p(\mathbf{b}_i)$ for them. This can be viewed as maximum a posterior (MAP) estimation, where the objective $p(\mathbf{X}|\mathbf{G},\mathbf{B})p(\mathbf{G})p(\mathbf{B})$ to be maximized consists of the observation model $p(\mathbf{X}|\mathbf{G},\mathbf{B})$ which measures the error of the approximation (1), and $p(\mathbf{G})$ and $p(\mathbf{B})$ are the priors of the gains and basis vectors, respectively. Here $\mathbf{X}$, $\mathbf{G}$, and $\mathbf{B}$ are matrices where all the observation vectors, gains, and basis vectors are grouped into their respective matrices. A generic formulation for the use of priors, adaptation, and MAP estimation in the separation is given in [5]. Instead of MAP estimation, full Bayesian treatment can also be used with appropriate priors [6].

The previous work [9] trained Gamma priors for the basis vectors since the Gamma distribution is a conjugate prior of the Poisson distribution which was used in the observation model. The work assumed a prior where the entries are statistically independent from each other, since this leads to computationally efficient algorithms. The assumption is unrealistic, since in natural audio spectra the frequencies are strongly dependent on each other. In this paper we propose an algorithm which allows modeling of the spectral covariance. The algorithm for training the priors is presented in Section 2, and the algorithm for performing the separation is presented in Section 3. In Section 4 the proposed method is shown to clearly improve the separation quality compared to previous methods where the covariance is not modeled.

2. TRAINING FULL-COVARIANCE PRIORS

The training is done using material where each source is present in isolation. In the previous work [9] we observed that it is beneficial to assume that each observation vector $\mathbf{s}_t^n$ of the training data is produced by a single basis vector. This means that in (1) only a single gain $g_{t,i}$ in each frame is non-zero. Training with the above assumption leads to basis vectors which correspond to entire spectra of target sources, instead of

parts of the spectra.

Prior distribution $p(\mathbf{B})$ of basis vectors can be efficiently learned by first normalizing the training data, and then estimating a mixture model for the normalized observations. The normalization effectively cancels out the gains so that the single non-zero gain becomes approximately unity. Each component in the mixture model corresponds to a single basis vector and models its distribution.

Observations $\mathbf{s}_t^n$ in our system are the square roots of energies measured on 80 frequency bands spaced uniformly on the mel scale (see Section 4 for details about the feature extraction). For natural sounds sources, the logarithms of energies measured in frequency bands are well modeled with normal distributions. However, frequency bands are usually strongly correlated. In many applications, discrete cosine transform is used to reduce the correlation. For example, mel-frequency cepstral coefficient features which are widely used in audio classification, are calculated this way.

Here we model the full covariance of basis vectors by training a full-covariance Gaussian mixture model for the log-spectrum features. First, log-spectral vectors of the training data are calculated and normalized so that the norm of each vector equals unity. The distribution of the normalized vectors $\mathbf{y}_t$ is modeled using a Gaussian mixture model

$$p(\mathbf{y}_t^n) = \sum_{i \in \mathcal{S}_n} w_i \mathcal{N}(\mathbf{y}_t^n; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{4}$$

where each Gaussian i has weight w_i, and $\mathcal{N}$ is a Gaussian distribution with mean vector $\boldsymbol{\mu}_i$ and full covariance matrix $\boldsymbol{\Sigma}_i$. The number of components in $\mathcal{S}_n$ is chosen beforehand and fixed.

The parameters are estimated by the expectation maximization (EM) algorithm, which is initialized by k-means clustering. Since the covariance matrices may become singular, the inverses of the covariance matrices are calculated by using the eigenvalue decomposition $\boldsymbol{\Sigma}_i = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\mathsf{T}$, where the eigenvalues in $\mathbf{V}$ are restricted above 10^{-6} times the largest eigenvalue. Furthermore, only 3 EM-algorithm iterations are used.

Each mixture component distribution is used as a prior for a log-basis vector, i.e.

$$p(\log(\mathbf{b}_i)) = \mathcal{N}(\log(\mathbf{b}_i); \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i). \tag{5}$$

Here $\log(\mathbf{b}_i)$ denotes the element-wise logarithm of vector $\mathbf{b}_i$. The basis vectors are assumed to be independent from each other. Figure 1 illustrates an example mean vector and covariance matrix. The mean of the distribution has peaky shape in the low frequencies, which means that it corresponds to a harmonic spectrum. In the covariance matrix, harmonic frequencies have relatively large strong correlations.

The algorithm can be made more efficient by decorrelating the dimensions of the normalized log-spectrum $\mathbf{y}_t^n$ vectors and reducing the number of dimensions by principal component analysis. Once the model (4) has been trained for the decorrelated vectors, the means and variances can be projected back to the original feature space.

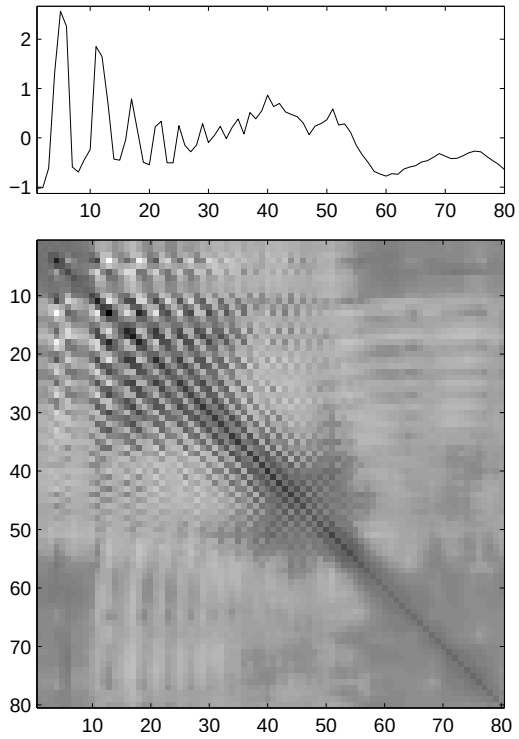

Figure 1: The mean vector (upper panel) and covariance matrix (lower panel) of an example log-basis vector. Dark color indicates positive correlation and light color negative correlation.

3. SEPARATION ALGORITHM

The separation algorithm maximizes the likelihood of the sources by maximizing the objective $p(\mathbf{X}|\mathbf{G},\mathbf{B})p(\mathbf{G})p(\mathbf{B})$. The negative logarithm of the objective is

$$L = D(\mathbf{X}||\mathbf{G},\mathbf{B}) + L_g(\mathbf{G}) + L_b(\mathbf{B}), \tag{6}$$

where the terms D, L_g, L_b account for the observation model $p(\mathbf{X}|\mathbf{G},\mathbf{B})$, prior of gains $p(\mathbf{G})$, and prior of basis vectors $p(\mathbf{B})$, respectively.

3.1 Objectives

We use a Poisson observation model, which negative logarithm is the divergence [9]. We use a gradient descent algorithm (explained later) in the estimation of the basis vectors, and the gradient of the divergence may have unbounded large terms, causing convergence problems. We circumvent the problem by adding a small constant ϵ to the model in the right side of Eq. (3). This can be viewed as a small noise floor in the model, and the resulting divergence is defined as

$$D(\mathbf{X}||\mathbf{G},\mathbf{B}) = \sum_{t=1}^{T}\sum_{f=1}^{F} d\left(x_{t,f}, \sum_{n=1}^{N}\sum_{i \in \mathcal{S}_n} b_{i,f} g_{t,i} + \epsilon\right),$$

where $d(p,q) = p\log(p/q) - p + q$. $x_{t,f}$ and $b_{i,f}$ denote the fth entry of vectors $\mathbf{x}_t$ and $\mathbf{b}_i$, respectively.

The model is slightly similar to the augmented divergence [8, pp. 34-48], but for simplicity ϵ is here added

only to the model, not to the observations. In our system ϵ had value which was 10^{-3} times the mean of the entries of $\mathbf{X}$. However, the method is not sensitive to the exact value of ϵ.

We use use an i.i.d. exponential distribution for the gains, and therefore the negative log of $p(\mathbf{G})$ (with the terms independent of the gains omitted) equals

$$L_g(\mathbf{G}) = \lambda \sum_{n=1}^{N} \sum_{i \in \mathcal{S}_n} \sum_{t=1}^{T} g_{t,i}, \tag{7}$$

where λ is the rate parameter of the distribution.

Term $p(\mathbf{B})$ corresponds to the full-covariance multivariate Gaussian distribution trained for each $\log(\mathbf{b}_i)$ in the training phase. We write the negative logarithm of the prior as a function of $\log(\mathbf{b}_i)$

$$L_b(\mathbf{B}) = \frac{1}{2} \sum_{n=1}^{N} \sum_{i \in \mathcal{S}_n} (\log(\mathbf{b}_i) - \boldsymbol{\mu}_i)^\mathsf{T} \boldsymbol{\Sigma}_i^{-1} (\log(\mathbf{b}_i) - \boldsymbol{\mu}_i),$$

where the terms independent of the basis vectors are omitted. The above is very similar to the term used by Wilson et al. in [10] to model the joint distribution of gains.

3.2 Iterative algorithm

The objective (6) is minimized using an iterative algorithm, where the gains and basis vectors are updated by turns. The gains are updated using multiplicative update

$$g_{t,i} \leftarrow g_{t,i} \frac{\mathbf{r}_t^\mathsf{T} \mathbf{b}_i}{\mathbf{1}^\mathsf{T} \mathbf{b}_i + \lambda}, \tag{8}$$

where

$$\mathbf{r}_t = \mathbf{x}_t ./ \left[\sum_{n=1}^{N} \sum_{i \in \mathcal{S}_n} \mathbf{b}_i g_{t,i} + \epsilon \right], \tag{9}$$

$./$ denotes element-wise division, and $\mathbf{1}$ is a all-one column vector.

Estimation of basis vectors is more problematic. Basically the algorithm proposed in [2] and used in [10] could be used here. The entries of the basis vector are heavily correlated, but multiplicative updates essentially update each element independently from each other. For a prior with strong covariance structure the update should be taken into the direction of the negative gradient.

We use gradient descent, which resulted to better convergence and results than the update proposed in [2]. The gradient of the divergence with respect to basis vector i is

$$\nabla_i D(\mathbf{X} || \mathbf{G}, \mathbf{B}) = \sum_t \mathbf{b}_i^\mathsf{T} \left(\mathbf{1} - \frac{\mathbf{x}_t}{\sum_{n=1}^{N} \sum_{i \in \mathcal{S}_n} \mathbf{b}_i g_{t,i} + \epsilon} \right),$$

where $\frac{\mathbf{x}}{\mathbf{y}}$ denotes element-wise division. The gradient of the negative log-prior with respect to basis vector i is

$$\nabla_i L_b(\mathbf{B}) = [\boldsymbol{\Sigma}_i^{-1} (\log(\mathbf{b}_i + \epsilon) - \boldsymbol{\mu}_i)] ./ (\mathbf{b}_i + \epsilon). \tag{10}$$

Also here we add small constant ϵ to ensure numerical stability.

The whole iterative estimation algorithm is given as follows.

1. Initialize each basis vector $\mathbf{b}_i$ with the distribution mean $\exp(\boldsymbol{\mu}_i)$. Initialize the gains with random positive values.
2. Calculate gradient $\nabla_i = \nabla_i D(\mathbf{X} || \mathbf{G}, \mathbf{B}) + \nabla_i L_b(\mathbf{B})$ with respect of each basis vector $\mathbf{b}_i$.
3. Update basis vectors into the negative direction of the gradient as $\mathbf{b}_i \leftarrow \mathbf{b}_i - \alpha \nabla_i$. Negative entries of the basis vectors are set of zero. Step size α is adapted by increasing it after iterations when the cost (6) decreased, and decreasing when the cost increased.
4. Update gains using multiplicative update rule (8).

The steps 2-4 are repeated until the algorithm converges. In practise we found it advantageous to use the moment term of the gradient, so that the effective gradient in Step 3 is the sum of the gradient of the current and previous iteration.

In total the computation time of the proposed method is a couple of times the computation time of the NMF algorithm [9] where independent Gamma distributions are used for each entry of $\mathbf{b}_i$, and multiplicative update rules were used for both the gains and basis vectors.

3.3 Source reconstruction

Under the Poisson model assumption, the expected value of the magnitude spectrum vector of each component is the ratio of its gain times the basis vector to the sum of all the gains times the basis vectors. [1, Section 2.1.1]. Therefore, the expected value of the spectrum vector produced by source n is given as

$$\hat{\mathbf{s}}_t^n = \mathbf{x}_t . * \frac{\sum_{i \in \mathcal{S}_n} g_{t,i} \mathbf{b}_i}{\sum_{n=1}^{N} \sum_{i \in \mathcal{S}_n} g_{t,i} \mathbf{b}_i + \epsilon}, \tag{11}$$

where $.*$ denotes element-wise multiplication.

4. SIMULATIONS

The proposed method was evaluated in separating signals consisting of a random male and a random female speaker. We use signals from the Grid corpus, [3], which consist of short sentences spoken by 34 speakers. 300 random test signals were generated. In each test signal, a random male speaker and a random female speaker were chosen. Three random sentences from both speakers were concatenated, and the speakers were mixed at equal power level.

Magnitude spectrum vectors were calculated as follows. First, the signal was filtered using a high-frequency emphasis filter. Then the signal was windowed into 32 ms frames with a Hamming window with 50 % overlap between adjacent frames. In each frame, the energy within 80 Mel-frequency bands was calculated. The algorithm operates on the square roots of the energies. The data representation is similar to the one used in [7] and [9].

In the training phase a model for both genders was trained. Each test speaker at time was excluded from the training data in order to simulate a situation where training data from a particular speaker is not available, but a gender model was in hand. This leave-one-out

training resulted in 18 male and 16 female models in total. Every 10th sentence (in the alphabetical order) of the training data was used to keep the computation time moderate.

4.1 Algorithms

The following algorithms were tested:

- FULL method is the proposed method where full-covariance priors and gradient descent are used to separate the sources.
- DIAG is otherwise similar to the proposed method, but diagonal covariance matrices are used in the priors.
- GAMMA method trains Gamma mixture models instead of Gaussian mixture models as explained in [9], and uses multiplicative updates to estimate parameters in the separation phase.
- SINGLE method allows only a single gain from each source to be non-zero in each frame. The parameters of this model were estimated by testing all the possible active component pairs, and selecting the one which resulted to the lowest divergence.
- SPEAKER method uses the proposed method, but the priors are trained for each test speaker. This studies a speaker-dependent separation setting where the test speakers are known in advance.

All the algorithms were tested with 30 and 70 basis vectors per speaker. In all the algorithms, the observation model and the prior were balanced by scaling the term L_b with a scaling factor 10 which resulted in approximately the best performance in the case of all the algorithms.

We also used sparseness factor λ which produced approximately the best results. Normalizing each basis vector to unity norm and scaling the distributions accordingly by multiplying the scale parameter was also found to improve the results slightly.

The SINGLE method evaluates the effect of the training phase assumption that a single component is active in each frame also in the separation phase. The above assumption is similar to the graphical models used in multi-talker speech recognition [4], where the combination of most likely state transition paths for the observation sequence is estimated.

4.2 Testing

Given a particular test signal, the prior sets learned for male and female speakers were combined, and the gains and basis vectors were estimated. All the algorithms except SINGLE were tested with fixed and adaptive basis vectors: fixed basis vectors used the means of the distributions without basis vector adaptation, whereas adaptive basis vectors denote methods where prior distributions are used.

The magnitude spectrum vectors of the male and female speaker in each frame were reconstructed according to Eq. (11). The quality of separation was measured by the signal-to-noise ratio (SNR) of the separated spectrograms. The SNRs were averaged over both the speakers in all the test signals.

Table 1: Signal-to-noise ratios of the tested methods in dB, obtained with fixed and adaptive basis vectors and with either 30 or 70 components per source. The best speaker-independent algorithm in each column is highlighted with bold face font.

	30 components		70 components	
method	fixed	adaptive	fixed	adaptive
GAMMA	6.55	6.73	6.95	7.04
FULL	**6.60**	**6.94**	**7.05**	**7.29**
DIAG	6.54	6.86	7.02	7.15
SINGLE	6.37	–	6.82	–
SPEAKER	6.74	7.01	6.75	6.92

4.3 Results

The average signal-to-noise ratios of each of the tested algorithm are illustrated in Table 1. In comparison with the other speaker-independent methods the proposed method produces better results. Even when fixed basis vectors are used the SNRs are slightly higher, but when adaptation is used the difference becomes more prominent.

With 30 components per source the speaker-dependent SPEAKER method produces better results than speaker-independent methods, but interestingly it performs worse when 70 components are used. The reason for this might be that there is not enough speaker-specific data to train reliably 70 basis vectors.

It is interesting to note that the SINGLE method resulted to slightly worse SNR than the proposed NMF methods, which means that modeling the observation as a weighted sum of basis vectors instead of a single basis vector is beneficial at least when the quality is measured by the SNR.

5. CONCLUSIONS

We have proposed a method to model the covariance of the spectrum in the prior distributions of the non-negative matrix factorization based sound source separation. Since the frequencies of natural sound sources are strongly correlated, we have to use an algorithm which is able to take into account the correlation. The gradient descent algorithm used in our system is able to produce good results while being computationally feasible. In comparison with the previous methods where frequency bins were assumed statistically independent from each other, the proposed covariance modeling technique leads to significantly better separation quality.

Acknowledgment

This work has been funded by the Academy of Finland. The author would like to thank Taylan Cemgil for his helpful comments.

REFERENCES

[1] A. T. Cemgil. Bayesian inference in non-negative matrix factorisation models. technical re-

port CUED/F-INFENG/TR.609. Technical report, University of Cambridge, July 2008.

[2] A. Cichocki, R. Zdunek, and S. Amari. New algorithms for non-negative matrix factorization in applications to blind source separation. In *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, Toulouse, France, 2006.

[3] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *Journal of the Acoustical Society of America*, 120(5), 2006.

[4] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson. Super-human multi-talker speech recognition: A graphical modeling approach. *Computer Speech and Language*, 2009. In press.

[5] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval. Adaptation of Bayesian models for single channel source separation and its application to voice / music separation in popular songs. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5), 2007.

[6] S. J. Rennie, J. R. Hershey, and P. A. Olsen. Efficient model-based speech separation and denoising using non-negative subspace analysis. In *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, Las Vegas, USA, 2008.

[7] M. N. Schmidt and R. K. Olsson. Single-channel speech separation using sparse non-negative matrix factorization. In *Proceedings of the International Conference on Spoken Language Processing*, Pittsburgh, USA, 2006.

[8] T. Virtanen. *Sound Source Separation in Monaural Music Signals*. PhD thesis, Tampere University of Technology, 2006. Available at `http://www.cs.tut.fi/~tuomasv`.

[9] T. Virtanen and A. T. Cemgil. Mixtures of gamma priors for non-negative matrix factorization based speech separation. In *Proceedings of the 8th International Conference on Independent Component Analysis and Blind Signal Separation*, 2009.

[10] K. W. Wilson, B. Raj, and P. Smaragdis. Regularized non-negative matrix factorization with temporal dependencies for speech denoising. In *9th Annual Conference of the International Speech Communication Association (Interspeech 2008)*, Brisbane, Australia, 2008.