Supplementary Material: Ambisonics domain Singing Voice Separation combining Deep Neural Network and Direction Aware Multichannel NMF

Antonio J. Muñoz-Montoro
Computer Science Department
Universidad de Oviedo
Gijón, Spain
munozantonio@uniovi.es

Julio J. Carabias-Orti, Pedro Vera-Candeas
Telecommunication Engineering Department
Universidad de Jaén
Jaén, Spain
carabias@ujaen.es, pvera@ujaen.es

I. Diagonalizable Itakura-Saito NTF variant

The Itakura Saito divergence between the observed $\mathbf{X}(f,t)$ and the estimated $\hat{\mathbf{X}}(f,t)$ SCM observations is expressed as:

$$\begin{aligned} D_{IS}(\mathbf{X}(f,t),\hat{\mathbf{X}}(f,t)) =& tr(\mathbf{X}(f,t)\hat{\mathbf{X}}(f,t)^{-1}) \\ & - log(det(\mathbf{X}(f,t)\hat{\mathbf{X}}(f,t)^{-1})) \\ & - K \end{aligned} \quad (1)$$

which omitting constant terms can be written as a function of the free parameters as follows:

$$f(\mathbf{Z},\mathbf{B},\mathbf{G}) = \Big[\mathrm{tr}(\mathbf{X}(f,t)\hat{\mathbf{X}}(f,t)^{-1}) - \log\det(\hat{\mathbf{X}}(f,t))\Big] \quad (2)$$

As explained in [1], to minimize this function, we follow the optimization scheme of majorization, in which an auxiliary (majorization) function is used. Let us define an auxiliary function f^{+}. In particular, the auxiliary function for the case of IS is defined as:

$$\begin{aligned} f^{+}(\mathbf{Z},\mathbf{B},\mathbf{G},\mathbf{R},\mathbf{U}) = & \\ \sum_{f,t}\Bigg[\log\det\mathbf{U}_{ft} + \frac{\det\mathbf{X}(f,t) - \det\mathbf{U}_{ft}}{\det\mathbf{U}_{ft}} & \\ + \sum_{s,o,k}\frac{\mathrm{tr}(\mathbf{X}(f,t)\mathbf{R}^{H}_{ftqdr}\mathbf{Y}_{d}^{-1}\mathbf{R}_{ftqdr})}{z_{qd}b_{qfr}g_{qrt}}\Bigg] & \end{aligned} \quad (3)$$

where auxiliary variables $\mathbf{R}_{ftqdr}$ and $\mathbf{U}_{ft}$ are hermitian positive definite matrices satisfying $\sum_{s,o,k}\mathbf{R}_{ftqdr} = [\mathrm{I}]^{K\times K}$ and $\mathbf{U}_{ft} = \mathbf{U}^{H}_{ft}$.

Then the derivation of the algorithm updates is achieved by minimizing the auxiliary function f^{+} w.r.t $\mathbf{R}_{ftqdr}$ and $\mathbf{U}_{ft}$ and then minimize via partial derivation w.r.t each model parameter and setting these derivatives to zero. See [1] for further details.

978-1-7281-9320-5/20/$31.00 ©2020 European Union

Although the IS divergence provide superior separation results in comparison with the squared frobenius norm based factorization. The low updating speed for such a model is mainly due to the inversion of a spatial covariance matrix, for which the complexity increases with the number of microphones,K, and is generally of order $O(K^3)$.

A way to mitigate the issue is using a diagonalization scheme [2]–[5]. The idea is to concentrate the energy on the diagonal part of the SCM to circumvent the matrix inversion, which can reduce the complexity to $O(K)$. Several full-rank methods uses a diagonalization matrix [2], [4]. Other methods applied a transformation matrix to both the observation and the estimation and then discard the off-diagonal values leading to a Non-negative Tensor Factorization (NTF) scheme [3], [5].

In this paper, we propose reduce to NTF by discarding the off-diagonal values in $\mathbf{X}(f,t)$, $\hat{\mathbf{X}}(f,t)$ and $\mathbf{Y}_d$. Consequently, the update rules in equations Eq. (12), Eq. (16) and Eq. (17) are replaced by:

$$z_{qd} \leftarrow z_{qd}\sqrt{\frac{\sum_{f,t}\hat{s}_q(f,t)\sum_k\frac{\mathrm{diag}(\mathbf{X}(f,t))}{\mathrm{diag}(\hat{\mathbf{X}}(f,t)^2)}\mathrm{diag}(\mathbf{Y}_d)}{\sum_{f,t}\hat{s}_q(f,t)\sum_k\frac{1}{\mathrm{diag}(\hat{\mathbf{X}}(f,t)^2)}\mathrm{diag}(\mathbf{Y}_d)}}. \quad (4)$$

$$b_{qfr} \leftarrow b_{qfr}\sqrt{\frac{\sum_{t,k}g_{qrt}\sum_k\frac{\mathrm{diag}(\mathbf{X}(f,t))}{\mathrm{diag}(\hat{\mathbf{X}}(f,t)^2)}\sum_d z_{qd}\,\mathrm{diag}(\mathbf{Y}_d)}{\sum_{t,k}g_{qrt}\sum_k\frac{1}{\mathrm{diag}(\hat{\mathbf{X}}(f,t)^2)}\sum_d z_{qd}\,\mathrm{diag}(\mathbf{Y}_d)}}. \quad (5)$$

$$g_{qrt} \leftarrow g_{qrt}\sqrt{\frac{\sum_{f,k}b_{qfr}\sum_k\frac{\mathrm{diag}(\mathbf{X}(f,t))}{\mathrm{diag}(\hat{\mathbf{X}}(f,t)^2)}\sum_d z_{qd}\,\mathrm{diag}(\mathbf{Y}_d)}{\sum_{f,k}b_{qfr}\sum_k\frac{1}{\mathrm{diag}(\hat{\mathbf{X}}(f,t)^2)}\sum_d z_{qd}\,\mathrm{diag}(\mathbf{Y}_d)}}. \quad (6)$$

where the function $\mathrm{diag}(\cdot)$ extracts the K diagonal elements of the SCM and reduces it to a 3-valence tensor.

REFERENCES

[1] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel Extensions of Non-Negative Matrix Factorization With Complex-Valued Data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 971–982, 5 2013.

[2] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Semi-supervised multichannel speech enhancement with a deep speech prior," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2197–2212, 2019.

[3] Y. Mitsufuji, S. Uhlich, N. Takamune, D. Kitamura, S. Koyama, and H. Saruwatari, "Multichannel Non-Negative Matrix Factorization Using Banded Spatial Covariance Matrices in Wavenumber Domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 49–60, 2020.

[4] N. Ito and T. Nakatani, "Fastmnmf: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 371–375, 2019.

[5] Y. Mitsufuji, N. Takamune, S. Koyama, and H. Saruwatari, "Multichannel blind source separation based on evanescent-region-aware non-negative tensor factorization in spherical harmonic domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 607–617, 2021.