SUPPLEMENTARY MATERIAL FOR THE IEEE WASPAA'17 SUBMISSION:

EXPLOITING THE INTERMITTENCY OF SPEECH FOR JOINT SEPARATION AND DIARISATION

Dionyssos Kounades-Bastian[1], *Laurent Girin*[1,2], *Xavier Alameda-Pineda*[1], *Radu Horaud*[1], *Sharon Gannot*[3]

[1] INRIA Grenoble Rhône-Alpes, France
[2] Univ. Grenoble Alpes, GIPSA-lab, France
[3] Bar-Ilan University, Faculty of Engineering, Israel

1. INTRODUCTION

This report is supplementary material for submission [1]. In [1] we want to recover the STFT coefficients $\left\{\mathbf{y}_{j,f\ell} \in \mathbb{C}^I\right\}_{j=1}^J$ of the J source images $\forall f, \ell$. Let $\mathbf{y}_{f\ell} = \left[\mathbf{y}_{1,f\ell}^\top \dots \mathbf{y}_{J,f\ell}^\top\right]^\top \in \mathbb{C}^{IJ}$ be the catenated vector of all J source images at time-frequency point f, ℓ.

1.0.1. Mixing Equation Revisited

Let the matrix $\mathbf{M}_n \in \mathbb{N}^{I \times IJ}$ be:

$$\mathbf{M}_n = \mathbf{d}_n^\top \otimes \mathbf{I}_I, \tag{1}$$

with $\otimes$ the Kronecker product. The observation $\mathbf{x}_{f\ell}$ equals the sum of active source-images plus some noise $\mathbf{b}_{f\ell} \in \mathbb{C}^I$:

$$\mathbf{x}_{f\ell} = \sum_{j=1}^J d_{j,Z_\ell} \mathbf{y}_{j,f\ell} + \mathbf{b}_{f\ell} = \tag{2}$$

$$\mathbf{M}_{Z_\ell} \mathbf{y}_{f\ell} + \mathbf{b}_{f\ell}. \tag{3}$$

Now let also $p(\mathbf{b}_{f\ell}) = \mathcal{N}_c\left(\mathbf{b}_{f\ell}; \mathbf{0}, o_f \mathbf{I}_I\right)$ and we obtain the observation model (eq. (4) in [1]): (parameters are omitted when denoting probabilities, that is $p(x;\theta)$ is simply denoted $p(x)$):

$$p\left(\mathbf{x}_{f\ell} \middle| Z_\ell = n, \mathbf{y}_{f\ell}\right) = \mathcal{N}_c\left(\mathbf{x}_{f\ell}; \mathbf{M}_n \mathbf{y}_{f\ell}, o_f \mathbf{I}_I\right). \tag{4}$$

The symbol $\mathcal{N}_c\left(\right)$ denotes the proper complex Gaussian distribution [2].[1]

[1] The proper complex Gaussian distribution is defined as $\mathcal{N}_c(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\pi\boldsymbol{\Sigma}|^{-1} \exp\left(-[\mathbf{x}-\boldsymbol{\mu}]^{\mathrm{H}} \boldsymbol{\Sigma}^{-1} [\mathbf{x}-\boldsymbol{\mu}]\right)$, with $\mathbf{x}, \boldsymbol{\mu} \in \mathbb{C}^I$ and $\boldsymbol{\Sigma} \in \mathbb{C}^{I \times I}$ being the argument, mean vector, and covariance matrix respectively.

1.0.2. The Prior Distribution of Source Images

As all J source images are a priori independent we can calculate the prior distribution of the catenated image $\mathbf{y}_{f\ell}$ with:

$$p(\mathbf{y}_{f\ell}) = \prod_{j=1}^J p(\mathbf{y}_{j,f\ell}) = \tag{5}$$

$$\prod_{j=1}^J \mathcal{N}_c\left(\mathbf{y}_{j,f\ell}; \mathbf{0}, u_{j,f\ell} \mathbf{R}_{j,f}\right) = \tag{6}$$

$$\mathcal{N}_c\left(\mathbf{y}_{f\ell}; \mathbf{0}_I, \operatorname{diag}_J\left(u_{j,f\ell} \mathbf{R}_{j,f}\right)\right), \tag{7}$$

with $\operatorname{diag}_J(\mathbf{A}_j)$ the $IJ \times IJ$ block-diagonal matrix with j-th diagonal block $\mathbf{A}_j$.

2. EM ALGORITHM

Fig. 1 shows the dependencies between hidden random variables and observations for the probabilistic model of [1]. Let $\mathbf{x}_{1:F1:L}$ be a short-hand for a set, i.e. $\{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L}$.

2.0.3. Complete Data Probability Distribution

The source images are assumed independent between all f, ℓ, j (as in [3, 4]), the observations $\mathbf{x}_{f\ell}$ are also independent over f, ℓ. Therefore, the completed data (observed and hidden variables) probability $p(\mathbf{y}_{1:F1:L}, Z_{1:L}, \mathbf{x}_{1:F1:L})$ for the model in [1] writes:

$$p(\mathbf{y}_{1:F1:L}, Z_{1:L}, \mathbf{x}_{1:F1:L}; \theta) =$$

$$p(Z_1) \prod_{\ell=2}^L p(Z_\ell | Z_{\ell-1}) \prod_{f,\ell=1}^{F,L} p(\mathbf{y}_{f\ell}) p(\mathbf{x}_{f\ell} | \mathbf{y}_{f\ell}, Z_\ell). \tag{8}$$

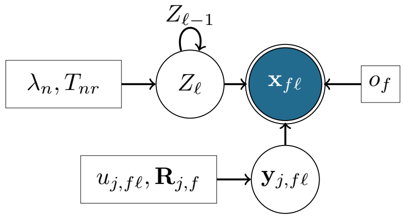

Figure 1: Associated graphical model: White circles denote hidden variables. Shaded (blue) circles denote observed variables. Loops denote temporal dependencies. Rectangles denote parameters to be estimated.

2.0.4. *Factorising the Posterior Distribution*

In the EM we want to derive the posterior distribution $p(\mathbf{y}_{1:F1:L}, Z_{1:L}|\mathbf{x}_{1:F1:L})$. From the Bayes rule we have:

$$p(\mathbf{y}_{1:F1:L}, Z_{1:L}|\mathbf{x}_{1:F1:L}) \propto \tag{9}$$
$$p(\mathbf{y}_{1:F1:L}, Z_{1:L}, \mathbf{x}_{1:F1:L}) \propto \tag{10}$$
$$p(\mathbf{y}_{1:F1:L}|Z_{1:L}, \mathbf{x}_{1:F1:L})p(Z_{1:L}|\mathbf{x}_{1:F1:L}). \tag{11}$$

Therefore replacing (11) on (8) we obtain:

$$p(\mathbf{y}_{1:F1:L}|Z_{1:L}, \mathbf{x}_{1:F1:L})p(Z_{1:L}|\mathbf{x}_{1:F1:L}) \propto$$
$$p(Z_1)\prod_{\ell=2}^{L} p(Z_\ell|Z_{\ell-1}) \prod_{f,\ell=1}^{F,L} p(\mathbf{y}_{f\ell})p(\mathbf{x}_{f\ell}|\mathbf{y}_{f\ell}, Z_\ell). \tag{12}$$

Therefore, isolating the terms from (12) that depend on $\mathbf{y}_{f\ell}$ yields its posterior $p(\mathbf{y}_{f\ell}|\mathbf{x}_{1:F1:L})$. Equivalently, isolating the terms from (12) that contain Z_ℓ provides its posterior $p(Z_\ell|\mathbf{x}_{1:F1:L})$.

Now, in Sec. 2.1 we compute $p(\mathbf{y}_{f\ell}|\mathbf{x}_{1:F1:L})$, and in Sec. 2.2 we compute $p(Z_\ell|\mathbf{x}_{1:F1:L})$.

2.1. E step Source Separation

The posterior of a source image $p(\mathbf{y}_{f\ell}|Z_\ell, \mathbf{x}_{1:F1:L})$ is found with (8), by dropping all terms of (8) that are independent of $\mathbf{y}_{f\ell}$. Then (8) writes:[2]

$$p(\mathbf{y}_{f\ell}|Z_\ell, \mathbf{x}_{1:F1:L}) \propto p(\mathbf{x}_{f\ell}|Z_\ell, \mathbf{y}_{f\ell})p(\mathbf{y}_{f\ell}) \propto \tag{13}$$
$$\mathcal{N}_c\left(\mathbf{y}_{f\ell}; \hat{\mathbf{y}}_{f\ell Z_\ell}, \mathbf{\Sigma}_{f\ell Z_\ell}\right). \tag{14}$$

The posterior covariance matrix $\mathbf{\Sigma}_{f\ell n} \in \mathbb{C}^{IJ\times IJ}$ and the posterior mean vector $\hat{\mathbf{y}}_{f\ell n} \in \mathbb{C}^{IJ}$ are respectively computed (for every $Z_\ell = n, n \in [1, N]$) with:

$$\mathbf{\Sigma}_{f\ell n} = \left[\operatorname{diag}_J\left(u_{j,f\ell}\mathbf{R}_{j,f}\right)^{-1} + \frac{\mathbf{M}_n^\top \mathbf{M}_n}{o_f}\right]^{-1}, \tag{15}$$

$$\hat{\mathbf{y}}_{f\ell n} = \mathbf{\Sigma}_{f\ell n}\mathbf{M}_n^\top \frac{\mathbf{x}_{f\ell}}{o_f}, \tag{16}$$

[2]We work in $\propto$ and therefore any term independent of $\mathbf{y}_{f\ell}$ is a constant for $p(\mathbf{y}_{f\ell}|\mathbf{x}_{1:F1:L})$ and can be dropped.

2.1.1. *Woodbury on the Posterior Covariance $\mathbf{\Sigma}_{j,f\ell n}$*

Applying Eq. (156) from [5] on (15) we have:

$$\mathbf{\Sigma}_{f\ell n} = \operatorname{diag}_J\left(u_{j,f\ell}\mathbf{R}_{j,f}\right) - \operatorname{diag}_J\left(u_{j,f\ell}\mathbf{R}_{j,f}\right) \times$$
$$\mathbf{M}_n\mathbf{V}_{f\ell n}^{-1}\mathbf{M}_n^\top \operatorname{diag}_J\left(u_{j,f\ell}\mathbf{R}_{j,f}\right), \tag{17}$$

with $\mathbf{V}_{f\ell n} \in \mathbb{C}^{I\times I}$ defined as

$$\mathbf{V}_{f\ell n} = \mathbf{M}_n^\top \operatorname{diag}_J\left(u_{j,f\ell}\mathbf{R}_{j,f}\right)\mathbf{M}_n = \tag{18}$$
$$\sum_{j=1}^{J} d_{j,n}u_{j,f\ell}\mathbf{R}_{j,f}. \tag{19}$$

2.1.2. *The Block Structure of $\mathbf{\Sigma}_{f\ell n}$*

From (17) we can now partition $\mathbf{\Sigma}_{f\ell n}$ in J^2, $I \times I$ blocks: $\{\mathbf{\Sigma}_{jr,f\ell n} \in \mathbb{C}^{I\times I}\}_{j,r=1}^{J,J}$. We are interested on the covariance matrix $\mathbf{\Sigma}_{j,f\ell n} \in \mathbb{C}^{I\times I}$ of a specific source image $\mathbf{y}_{j,f\ell}$. that is the j-th, $I \times I$ diagonal block $\mathbf{\Sigma}_{jj,f\ell n}$:

$$\mathbf{\Sigma}_{jj,f\ell n} = u_{j,f\ell}\mathbf{R}_{j,f} -$$
$$d_{j,n}u_{j,f\ell}\mathbf{R}_{j,f}\mathbf{V}_{f\ell n}^{-1}d_{j,n}u_{j,f\ell}\mathbf{R}_{j,f}, \tag{20}$$

Eq. (20) corresponds to (10) in [1].

We will also need the non-diagonal blocks $\mathbf{\Sigma}_{jr,f\ell n}, j \neq r$ that are expressible with:

$$\mathbf{\Sigma}_{jr,f\ell n} = -d_{j,n}u_{j,f\ell}\mathbf{R}_{j,f}\mathbf{V}_{f\ell n}^{-1}d_{r,n}u_{r,f\ell}\mathbf{R}_{r,f}. \tag{21}$$

2.1.3. *The Posterior Mean $\hat{\mathbf{y}}_{j,f\ell n}$ of a Source Image*

We are interested on the posterior mean $\hat{\mathbf{y}}_{j,f\ell n} \in \mathbb{C}^I$ of a specific source image $\mathbf{y}_{j,f\ell}$, obtained from the respective part of the long vector $\hat{\mathbf{y}}_{f\ell n}$ that has been computed with (16).

We can simplify (16) by applying (158) from [5]:

$$\hat{\mathbf{y}}_{f\ell n} = \operatorname{diag}_J\left(u_{j,f\ell}\mathbf{R}_{j,f}\right)\mathbf{M}_n^\top\mathbf{V}_{f\ell n}^{-1}\mathbf{x}_{f\ell}. \tag{22}$$

Or simply for a specific $\hat{\mathbf{y}}_{j,f\ell} \in \mathbb{C}^I$:

$$\hat{\mathbf{y}}_{j,f\ell n} = u_{j,f\ell}\mathbf{R}_{j,f}d_{j,n}\mathbf{V}_{f\ell n}^{-1}\mathbf{x}_{f\ell}. \tag{23}$$

Clearly, (23) is equivalent with (9) in [1].

2.2. E step Source Diarisation

We compute $p(Z_{1:L}|\mathbf{x}_{1:F1:L})$ from (8), by marginalising out all source images:

$$p(Z_{1:L}|\mathbf{x}_{1:F,1:L}) = p(Z_1)\prod_{\ell=2}^{L} p(Z_\ell|Z_{\ell-1})\times$$

$$\prod_{f,\ell=1}^{F,L}\int_{\mathbf{y}_{f\ell}} p(\mathbf{x}_{f\ell}|Z_\ell,\mathbf{y}_{f\ell})p(\mathbf{y}_{f\ell})\mathrm{d}\mathbf{y}_{f\ell} = \tag{24}$$

$$p(Z_1)\prod_{\ell=2}^{L} p(Z_\ell|Z_{\ell-1})\times$$

$$\prod_{f,\ell=1}^{F,L}\mathcal{N}_c\left(\mathbf{x}_{f\ell};\mathbf{0},\mathbf{V}_{f\ell Z_\ell}\right). \tag{25}$$

where (for each $Z_\ell = n$) $\mathbf{V}_{f\ell n}$ is calculated with (19). As for the integral is calculated with Eq.(2.115) from [6].

2.2.1. Forward-Backward Algorithm for HMM

Eq. (25) is the joint distribution of an HMM with hidden state Z_ℓ along $\ell \in [1, L]$ (see Eq. (13.10) in [6]). and some emission probabilities $\iota_{\ell Z_\ell}$ defined:

$$\iota_{\ell Z_\ell} = \prod_{f=1}^{F}\mathcal{N}_c\left(\mathbf{x}_{f\ell};\mathbf{0},\mathbf{V}_{f\ell Z_\ell}\right). \tag{26}$$

The posterior probability $\eta_{\ell n} = p(Z_\ell = n|\mathbf{x}_{1:F1:L})$ of each hidden state is hence computed using the forward-backward algorithm: provided in equations (13.36), (13.38) of [6].

2.3. M step

In the M step, the parameters θ are updated by maximising the Expected Complete Data Log-likelihood (ECDLL) function (see Eq. (9.30) in [6]) with respect to the parameters θ.

2.3.1. M-T_{nr}, λ_n

The update rules for the diarisation parameters T_{nr}, λ_n are the ML updates for HMM parameters: Equations (13.19), (13.18) of [6].

2.3.2. M-$w_{j,fk}, h_{j,k\ell}, \mathbf{R}_{j,f}$

The source image parameters $w_{j,fk}, h_{j,k\ell}, \mathbf{R}_{j,f}\forall f,\ell,j$ are updated as in [4]. To apply the rules derived in [4] one needs the second order posterior moment of a source image $\mathbf{y}_{j,f\ell}$ that is found with:

$$\mathbf{Q}_{j,f\ell} = \sum_{n=1}^{N}\eta_{\ell n}\int_{\mathbf{y}_{f\ell}} p(\mathbf{y}_{f\ell}|Z_\ell = n,\mathbf{x}_{1:F1:L})\times$$

$$\mathbf{y}_{j,f\ell}\mathbf{y}_{j,f\ell}^{\mathrm{H}}\mathrm{d}\mathbf{y}_{f\ell} = \tag{27}$$

$$\sum_{n=1}^{N}\eta_{\ell n}\left(\mathbf{\Sigma}_{jj,f\ell n} + \hat{\mathbf{y}}_{j,f\ell n}\hat{\mathbf{y}}_{j,f\ell n}^{\mathrm{H}}\right). \tag{28}$$

2.3.3. M-o_f

The ECDLL $\mathcal{L}(o_f)$ regarding o_f writes:

$$\mathcal{L}(o_f) = \sum_{n=1}^{N}\eta_{\ell n}\int_{\mathbf{y}_{f\ell}} p(\mathbf{y}_{f\ell}|Z_\ell = n,\mathbf{x}_{1:F1:L})\times$$

$$\log\mathcal{N}_c\left(\mathbf{x}_{f\ell};\mathbf{M}_n\mathbf{y}_{f\ell},o_f\mathbf{I}_I\right)\mathrm{d}\mathbf{y}_{f\ell}. \tag{29}$$

Differentiating $\mathcal{L}(o_f)$ w.r.t. o_f and setting the result to zero yields the update rule for o_f:

$$o_f = \frac{1}{LI}\sum_{\ell=1}^{L}\Bigg(\mathbf{x}_{f\ell}^{\mathrm{H}}\mathbf{x}_{f\ell} - \left(\sum_{n=1}^{N}\eta_{\ell n}\hat{\mathbf{x}}_{f\ell n}\right)^{\mathrm{H}}\mathbf{x}_{f\ell} - \mathbf{x}_{f\ell}^{\mathrm{H}}\left(\sum_{n=1}^{N}\eta_{\ell n}\hat{\mathbf{x}}_{f\ell n}\right) + \sum_{n=1}^{N}\eta_{\ell n}\mathrm{tr}\Big\{\mathbf{M}_n\left(\mathbf{\Sigma}_{f\ell n} + \hat{\mathbf{y}}_{f\ell n}\hat{\mathbf{y}}_{f\ell n}^{\mathrm{H}}\right)\mathbf{M}_n^{\top}\Big\}\Bigg). \tag{30}$$

with $\hat{\mathbf{x}}_{f\ell n}$ defined as:

$$\hat{\mathbf{x}}_{f\ell n} = \mathbf{M}_n\hat{\mathbf{y}}_{f\ell n} = \sum_{j=1}^{J}\cancel{d_{j,n}}\hat{\mathbf{y}}_{j,f\ell n}. \tag{31}$$

Notice that $d_{j,n}$ is already applied on (23) and it does not need to be re-applied as it is binary.

2.3.4. SImplification of the Quadratic Term

Now let's work with the quadratic term in (30):

$$\mathrm{tr}\Big\{\mathbf{M}_n\left(\mathbf{\Sigma}_{f\ell n} + \hat{\mathbf{y}}_{f\ell n}\hat{\mathbf{y}}_{f\ell n}^{\mathrm{H}}\right)\mathbf{M}_n^{\top}\Big\} = \tag{32}$$

$$\mathrm{tr}\Big\{\mathbf{M}_n\mathbf{\Sigma}_{f\ell n}\mathbf{M}_n^{\top}\Big\} + \hat{\mathbf{x}}_{f\ell n}^{\mathrm{H}}\hat{\mathbf{x}}_{f\ell n}. \tag{33}$$

Now let us define the variance part of the above as $\delta_{f\ell n}$, which is practically the sum of all J^2 blocks of the source covariance that due to $\mathbf{M}_n$ are multiplied with the diarisation:

$$\delta_{f\ell n} = \mathrm{tr}\Big\{\mathbf{M}_n\mathbf{\Sigma}_{f\ell n}\mathbf{M}_n^{\top}\Big\} = \tag{34}$$

$$\mathrm{tr}\Big\{\sum_{j=1}^{J}\sum_{r=1}^{J} d_{j,n}d_{r,n}\mathbf{\Sigma}_{jr,f\ell n}\Big\} = \tag{35}$$

$$\mathrm{tr}\left\{\mathbf{P}_{f\ell n} - \mathbf{P}_{f\ell n}\mathbf{V}_{f\ell n}^{-1}\mathbf{P}_{f\ell n}\right\}. \tag{36}$$

3. REFERENCES

[1] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "Exploiting the intermittency of speech for joint separation and diarisation of speech signals," in *Submitted to IEEE Workshop Applicat. Signal Process. to Audio and Acoust. (WASPAA)*, New Paltz, NY, 2017.

[2] F. Neeser and J. Massey, "Proper complex random processes with applications to information theory," *IEEE Trans. Info. Theory*, vol. 39, no. 4, pp. 1293–1302, 1993.

[3] N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[4] S. Arberet, A. Ozerov, N. Q. K. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *IEEE Int. Conf. Info. Sciences, Signal Process., Applicat.*, Kuala Lumpur, Malaysia, 2010.

[5] K. B. Petersen and M. S. Pedersen, *The matrix cookbook*. Version. Nov 15, 2012.

[6] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.