Joint Audio Source Separation and Diarization

J Source signals, reverberated and summed, are being recorded at I microphones.

- Separation: Recover the J orginal source signals!
- Diarization: Classify each source as active/inactive along time!

Standard Formulation in the STFT domain

- Separate a mixture of J sources with I microphones.
- In STFT domain the problem becomes:

$$\mathbf{x}_{f\ell} = \mathbf{A}_f \mathbf{s}_{f\ell} + \mathbf{b}_{f\ell}$$

mixture $[I \times 1]$
observed

mixing matrix $[I \times J]$
unknown!

source STFT $[J \times 1]$
unknown!

sensor noise $[I \times 1]$
unknown!

- f=1:F frequency bins, ℓ=1:L time frames.

Modelling Diarization

- The Standard Model: $\mathbf{x}_{f\ell} = \mathbf{A}_f \mathbf{s}_{f\ell} + \mathbf{b}_{f\ell}$ has all sources active;
- Look between $\mathbf{A}_f$ and $\mathbf{s}_{f\ell}$: You notice the identity matrix????

$$\mathbf{x}_{f\ell} = \mathbf{A}_f \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} \begin{bmatrix} s_{1,f\ell} \\ s_{2,f\ell} \\ s_{3,f\ell} \end{bmatrix} + \mathbf{b}_{f\ell}.$$

- This is a special case where the diagonal entries are fixed to 1.
- What if an entry was 0 instead, e.g.

$$\mathbf{x}_{f\ell} = \mathbf{A}_f \begin{bmatrix} 1 & & \\ & 1 & \\ & & 0 \end{bmatrix} \begin{bmatrix} s_{1,f\ell} \\ s_{2,f\ell} \\ s_{3,f\ell} \end{bmatrix} + \mathbf{b}_{f\ell} =$$

$$\mathbf{a}_{1,f} s_{1,f\ell} + \mathbf{a}_{2,f} s_{2,f\ell} + \cancel{\mathbf{a}_{3,f} 0 s_{3,f\ell}},$$

- where $\mathbf{a}_{1,f}, .., \mathbf{a}_{3,f} \in \mathbb{C}^I$ are the columns of $\mathbf{A}_f$.

The state variable

- For $J = 3$ sources there are $N = 8$ possible matrices:

$$\mathbf{D}_1 = \begin{bmatrix} 0 & & \\ & 0 & \\ & & 0 \end{bmatrix}, \mathbf{D}_2 = \begin{bmatrix} 0 & & \\ & 0 & \\ & & 1 \end{bmatrix}, \ldots, \mathbf{D}_8 = \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix}.$$

- Let a categoric variable $Z_\ell = n, n \in [1, N]$ choose the $\mathbf{D}_n$ at time frame ℓ.
- The hidden variable Z_ℓ has a temporal model on $\ell = 1, .., L$

$$p(Z_\ell = n | Z_{\ell-1} = r) = T_{nr}, \quad p(Z_\ell = n) = \lambda_n,$$

with λ_n, T_{nr} prior and transition parameters.

Audio Mixture with Diarization

- We can compactify:

$$\mathbf{x}_{f\ell} = \mathbf{A}_f \mathbf{D}_{Z_\ell} \mathbf{s}_{f\ell} + \mathbf{b}_{f\ell},$$

- or probabilistically (white isotropic $\mathbf{b}_{f\ell}$):

$$p(\mathbf{x}_{f\ell} | Z_\ell = n, \mathbf{s}_{f\ell}) = \mathcal{N}_c \left(\mathbf{A}_f \mathbf{D}_n \mathbf{s}_{f\ell},\ \mathrm{v}_f \mathbf{I}_I \right),$$

where $\mathbf{A}_f \in \mathbb{C}^{I \times J}$ mixing matrix, and $\mathrm{v}_f \in \mathbb{R}_+$ microphone noise variance, parameters to be estimated.

Modelling the Sources (NMF)

- The prior distribution for $\mathbf{s}_{f\ell} \in \mathbb{C}^J$ is

$$p\left(\mathbf{s}_{f\ell}\right) = \mathcal{N}_c\left(\begin{bmatrix} 0 \\ \\ 0 \\ \\ 0 \end{bmatrix}, \begin{bmatrix} u_{1,f\ell} & & & & \\ & \ddots & & & \\ & & u_{j,f\ell} & & \\ & & & \ddots & \\ & & & & u_{J,f\ell} \end{bmatrix} \right),$$

- $u_{j,f\ell}$ are the source PSD (non-negative parameters).
- The $F \times L$ matrix $\left\{u_{j,f\ell}\right\}_{f,\ell=1}^{F,L}$ is factorised via Nonnegative Matrix Factorisation to reduce the number of parameters.

Associated Graphical Model

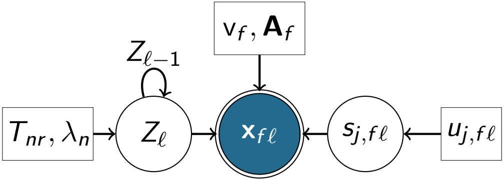


Figure: Joint Audio Source Separation and Diarization.

Associated EM algorithm

- Hidden variables:

$$\mathcal{H} = \{\mathbf{s}_{f\ell}, Z_\ell\}_{f,\ell=1}^{F,L}.$$

- Observed data:

$$\mathcal{X} = \{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L}.$$

- Joint a posteriori distribution **it factorizes!!!!!!!!!!!!!!!!**

$$p(\mathbf{s}_{1:F1:L}, Z_{1:L}|\mathcal{X}) \propto p(\mathbf{s}_{1:F1:L}|Z_{1:L}, \mathcal{X})p(Z_{1:L}|\mathcal{X}).$$

- Parameters to be estimated (e.g. $J = 3, N = 8$):

$$\theta = \left\{\mathbf{A}_f, \mathrm{v}_f, u_{j,f\ell}, T_{nr}, \lambda_n\right\}_{f,\ell,j,n,r=1}^{F,L,J,N,N}.$$

E-Step Outline 1/2

- E-step of **Sources|Diarization:** For every value of $Z_\ell = n$:

$$p(\mathbf{s}_{f\ell}|Z_\ell = n) = \mathcal{N}_c\left(\boldsymbol{\mu}^s_{f\ell n}, \boldsymbol{\Sigma}^s_{f\ell n}\right).$$

- Closed form expressions

$$\boldsymbol{\Sigma}^s_{f\ell n} = \left[\mathrm{diag}_J\left(\frac{1}{u_{j,f\ell}}\right) + \mathbf{D}_n \frac{\mathbf{A}_f^{\mathrm{H}}\mathbf{A}_f}{\mathrm{v}_f}\mathbf{D}_n\right]^{-1},$$

$$\boldsymbol{\mu}^s_{f\ell n} = \boldsymbol{\Sigma}^s_{f\ell n}\mathbf{D}_n\mathbf{A}_f^{\mathrm{H}}\frac{\mathbf{x}_{f\ell}}{\mathrm{v}_f}.$$

E-Step Outline 2/2

- E Step of **Diarization:** Estimate the Responsibilities of Z_ℓ:
- $\eta_{\ell,n} = p(Z_\ell = n|\mathcal{X})$ is found by solving a HMM with emission probability

$$p(\mathcal{X}|Z_\ell = n) \propto \exp\left(\sum_{f=1}^{F}\left(\log|\boldsymbol{\Sigma}^s_{f\ell n}| + \frac{\mathbf{x}^{\mathrm{H}}_{f\ell}}{\mathsf{v}_f}\mathbf{A}_f\mathbf{D}_n\boldsymbol{\mu}^s_{f\ell n}\right)\right),$$

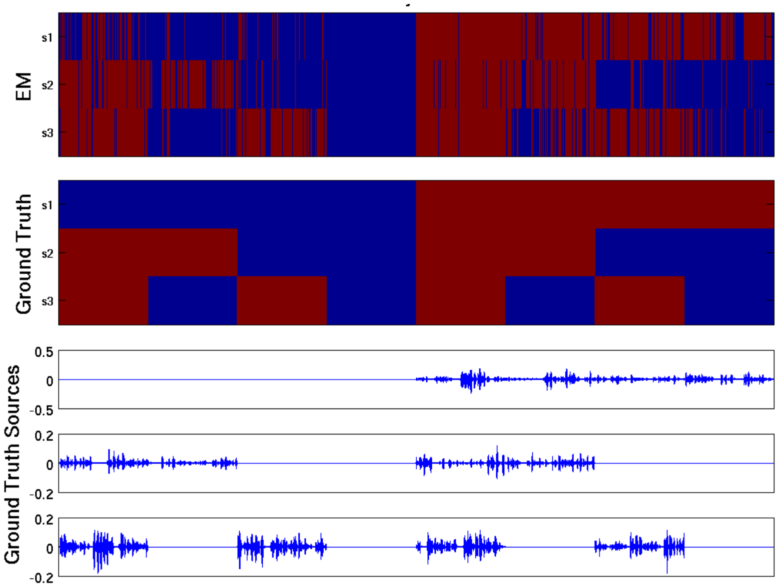
transition T_{nr}, and beginning-of-time probabilities λ_n.

- Closed form solution via the forward-backward algorithm.

M-step Outline

- The parameters $\mathbf{A}_f, \mathrm{v}_f$ updated in closed form: Typical rules for the Gaussian.
- T_{nr}, λ_n updated in closed form: provided for free by the forward-backward algorithm.
- $u_{j,f\ell}$ is composed by two sets of parameters $\{w_{fk}\}_{f,k}, \{h_{k\ell}\}_{k,\ell}$: Typical (closed form, alternating) updates for the NMF.

Estimated Activity on a Mix of $J = 3$ Sources via EM



Some Quantitative (Continiously active)

Comparisson of Separation performance (dB) on a 2×3 mix of continously emmiting sources:

	Proposed			Ozerov & Févotte '10		
	SDR	SIR	SAR	SDR	SIR	SAR
$s1$	9.2	13.4	13.6	9.3	13.8	14.0
$s2$	7.1	15.2	13.4	7.1	14.5	14.1
$s3$	9.6	14.0	13.9	9.6	13.4	14.6

Thank you !