Exploiting the Intermittency of Speech for Joint Separation and Diarisation of Speech Signals

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

[1]INRIA Grenoble Rhône-Alpes, [2]Grenoble INP, [3]Bar-Ilan University

Motivation

Speech signals are intermittent in natural conversations. Would sound source separation and diarization benefit from joint modeling?

Contribution

A probabilistic model and the correpsonding exact EM algorithm for joint source separation and diarization. We use a HMM for speaker activity modeling and the full-rank spatial covariance matrix model [1] with NMF for the sources.

Problem

Input

Traditionally, the mixture signal $x_i(t)$ at microphone $i \in [1, I]$ is the sum of J speech source images $y_{i,j}(t), j \in [1, J]$ plus the microphone noise $b_i(t)$:

$$x_i(t) = \sum_{j=1}^{J} y_{i,j}(t) + b_i(t).$$

Goal

To recover the J speech source images $y_{i,j}(t), j \in [1, J]$ and their activity, i.e. the *speaker diarization*. Thus we introduce a **hidden diarization variable** in the formulation.

Modeling Diarization

In the STFT domain and in vector form:

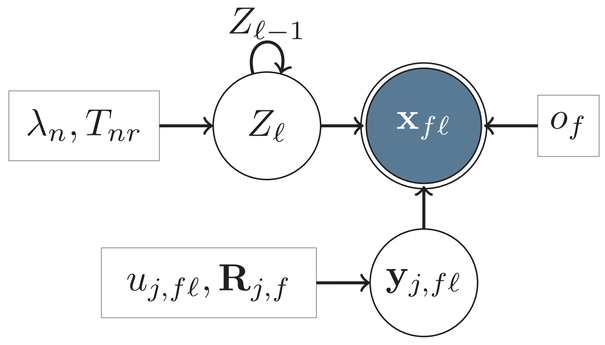
$$\mathbf{x}_{f\ell} = \sum_{j=1}^{J} d_{j,Z_\ell} \mathbf{y}_{j,f\ell} + \mathbf{b}_{f\ell} \in \mathbb{C}^I,$$

where d_{j,Z_ℓ} is a binary value indicating **the activity of the j-th source at frame ℓ** and Z_ℓ is a categorical variable taking values within $[1, N = 2^J]$. Example with $J = 2$

$$\mathbf{d}_1 = \left[\begin{smallmatrix}0\\0\end{smallmatrix}\right],\ \mathbf{d}_2 = \left[\begin{smallmatrix}0\\1\end{smallmatrix}\right],\ \mathbf{d}_3 = \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right],\ \mathbf{d}_4 = \left[\begin{smallmatrix}1\\1\end{smallmatrix}\right].$$

$d_{j,n}$ is j-th entry of $\mathbf{d}_n$.

Probabilistic Model



HMM for diarization

Z_ℓ is a hidden variable assumed to follow a first-order Markov chain:

$$p(Z_\ell = n | Z_{\ell-1} = r) = T_{nr}.\ \ p(Z_1 = n) = \lambda_n,$$

with λ_n, T_{nr} parameters and $n, r \in [1, N]$.

Observation model

$$p(\mathbf{x}_{f\ell} | Z_\ell = n) = \mathcal{N}_c\left(\mathbf{x}_{f\ell}; \sum_{n=1}^{N} d_{j,n} \mathbf{y}_{j,f\ell}, o_f \mathbf{I}\right).$$

Source model from [16]

$$p(\mathbf{y}_{j,f\ell}) = \mathcal{N}_c\left(\mathbf{y}_{j,f\ell}; \mathbf{0}, u_{j,f\ell}\mathbf{R}_{j,f}\right)$$

with $\mathbf{R}_{j,f}$ being the spatial covariance matrix and $u_{j,f\ell}$ being the source PSD:

$$u_{j,f\ell} = \sum_{j=1}^{J} w_{j,fk} h_{j,k\ell},$$

with non-negative $w_{j,fk}, h_{j,k\ell}$.

EM Algorithm

EM yields a source image estimate for each diarisation state $Z_\ell = n$:

$$\hat{\mathbf{y}}_{j,f\ell n} = \mathbf{G}_{j,f\ell n} \mathbf{V}_{f\ell n}^{-1} \mathbf{x}_{f\ell}$$

with $\mathbf{V}_{f\ell n}$ the covariance matrix of the mixture signal:

$$\mathbf{V}_{f\ell n} = \sum_{j=1}^{J} \mathbf{G}_{j,f\ell n} + o_f \mathbf{I}_I,$$

$$\mathbf{G}_{j,f\ell n} = d_{j,n} u_{j,f\ell} \mathbf{R}_{j,f}.$$

The posterior probability of diarization $\eta_{\ell n}$:

$$\eta_{\ell n} = p(Z_\ell = n | \{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L}),$$

is obtained using the forward-backward algorithm:

$$\phi_{\ell n} \propto \iota_{\ell n} \sum_{r=1}^{N} T_{nr} \phi_{(\ell-1)r},$$

$$\beta_{\ell n} \propto \sum_{r=1}^{N} T_{rn} \iota_{(\ell+1)r} \beta_{(\ell+1)r},$$

$$\eta_{\ell n} \propto \phi_{\ell n} \beta_{\ell n},$$

with $\iota_{\ell n} = \prod_{f=1}^{F} \mathcal{N}_c\left(\mathbf{x}_{f\ell}; \mathbf{0}, \mathbf{V}_{f\ell n}\right)$ the observation probability for diarization state $Z_\ell = n$. **The final source image estimate writes:**

$$\hat{\mathbf{y}}_{j,f\ell} = \sum_{n=1}^{N} \eta_{\ell n} \hat{\mathbf{y}}_{j,f\ell n}.$$

Results

Experimental setup: underdetermined stereo ($I = 2$) mixtures of $J = 3$ sources from TIMIT, BRIRs with $RT_{60} = 0,21$s.

Separation results

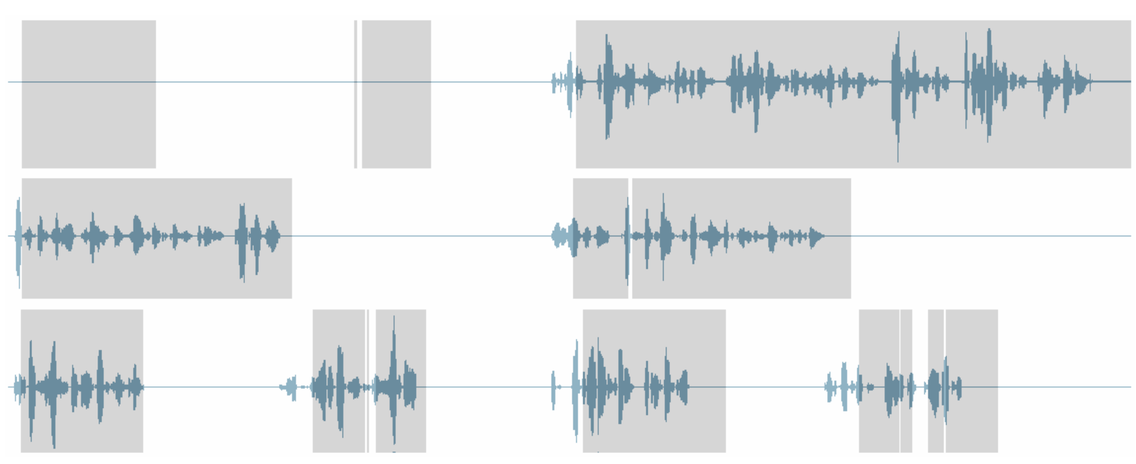
Speaker overlap	Metric (dB)	[24]	[25]	[17]	[1]	[16]	Prop.
Full	SDR	2.9	2.5	2.6	2.7	2.9	**3.4**
	SIR	4.9	5.2	6.0	5.3	5.6	**6.9**
	SAR	7.3	**8.3**	8.2	6.4	7.6	6.7
Partial	SDR	3.2	2.6	2.5	2.9	3.3	**4.2**
	SIR	5.2	5.2	5.5	5.6	6.1	**8.4**
	SAR	8.3	**9.3**	9.2	7.4	8.4	7.9
None	SDR	3.3	2.6	2.7	3.1	3.5	**5.0**
	SIR	5.9	5.8	6.4	6.4	7.1	**10.8**
	SAR	9.2	10.4	**10.5**	8.3	9.5	9.6

Diarization results

Speaker overlap	Method [8]	[14]	Prop.
Full	33.3	**92.2**	87.5
Partial	60.5	59.2	**70.0**
No	67.5	56.1	**69.5**

Diarization Accuracy is the % of STFT frames where the source activity is correctly detected.

Example of estimated diarization



Shade indicates intervals where the proposed method detected the source as active.

Conclusion

Improving speech separation and diarization performance by a joint formulation.

References

[1] N. Duong et. al., TASLP, 2010
[8] D. Vijayasenan et. al., Speech Process., 2012
[14] D.Kounades-Bastian et. al., ICASSP, 2017
[16] S. Arberet et. al., ISSPA, 2010
[24] Y. Dorfan et. al., TASLP, 2015
[25] A. Ozerov et. al., TASLP, 2010.