A source/filter model with adaptive constraints for NMF-based speech separation

Damien Bouvier[1], Nicolas Obin[1], Axel Roebel[1], Marco Liuni[1]

[1]IRCAM, UPMC - Sorbonne Universités, CNRS

International Conference on Acoustics, Speech and Signal Processing
24 March 2016

State of the art

- Speech separation using NMF
- Semi-supervised NMF
- Source/filter model

Proposed method

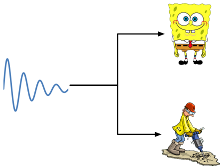
- Semi-supervised constrained NMF
- Contribution 1: speech-specific source/filter coherence constraint
- Contribution 2: adaptive weight method

Experimental evaluation

- Experiment description
- Effect of weight's adaptation
- Algorithm comparison

Speech separation using NMF

Signal has only 2 sources:
speech and background sound



Supervised algorithms

- [Mysore and Smaragdis, 2012]: language model
- [Virtanen et al., 2013]: new updates using Newton algorithm
- [Sun and Mysore, 2013]: Universal Speech Model (USM)

Semi-supervised algorithms

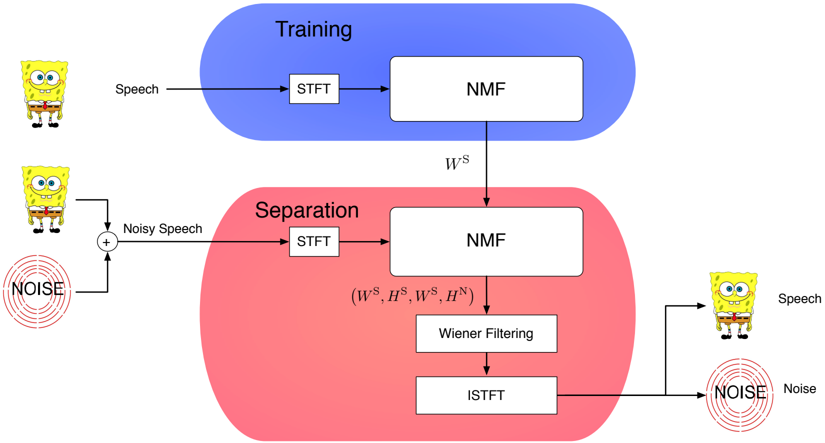
- [Germain and Mysore, 2015]: USM & online noise adaptation

Unspervised algorithms (but informed)

- [Le Magoarou et al., 2014]: use of textual information
- [Durrieu et al., 2009]: source/filter model for NMF

Semi-supervised NMF

$$\underset{\boldsymbol{W}^S,\boldsymbol{W}^N,\boldsymbol{H}^N \geq 0}{\operatorname{argmin}} \quad \mathcal{C}(\boldsymbol{V}|\widetilde{\boldsymbol{V}}) \text{ with } \begin{cases} \widetilde{\boldsymbol{V}} = \boldsymbol{W}^S\boldsymbol{H}^S + \boldsymbol{W}^N\boldsymbol{H}^N \\ \boldsymbol{W}^S \text{ learned} \end{cases} \tag{1}$$



Source/filter model [Durrieu et al., 2009]

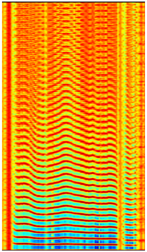
$$\text{source} \quad * \quad \text{filter} \quad = \quad \text{sound}$$

$$\text{glottis} \longrightarrow \boxed{\text{vocal tract}} \longrightarrow \text{speech}$$

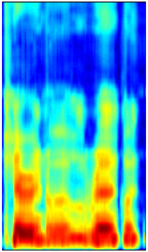
Source/filter model [Durrieu et al., 2009]

source $*$ filter $=$ sound

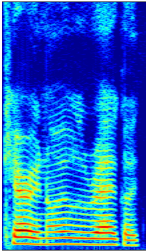
glottis $\longrightarrow$ $\boxed{\text{vocal tract}}$ $\longrightarrow$ speech



$.*$



$=$



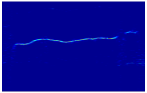
$$\boldsymbol{V}^{\text{ex}} \;.*\; \boldsymbol{V}^{\Phi} = \boldsymbol{V}$$

Source/filter model [Durrieu et al., 2009]

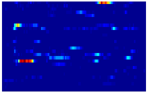
source * filter = sound

glottis ⟶ $\boxed{\text{vocal tract}}$ ⟶ speech

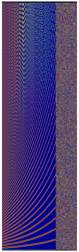
$\boldsymbol{H}^{\text{ex}}$

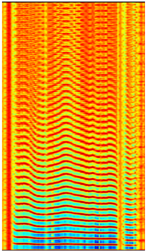


$\boldsymbol{H}^{\Phi}$



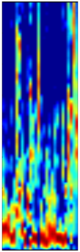
$\boldsymbol{W}^{\text{ex}}$

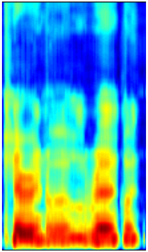




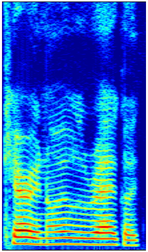
.*

$\widehat{\boldsymbol{W}}^{\Phi}$



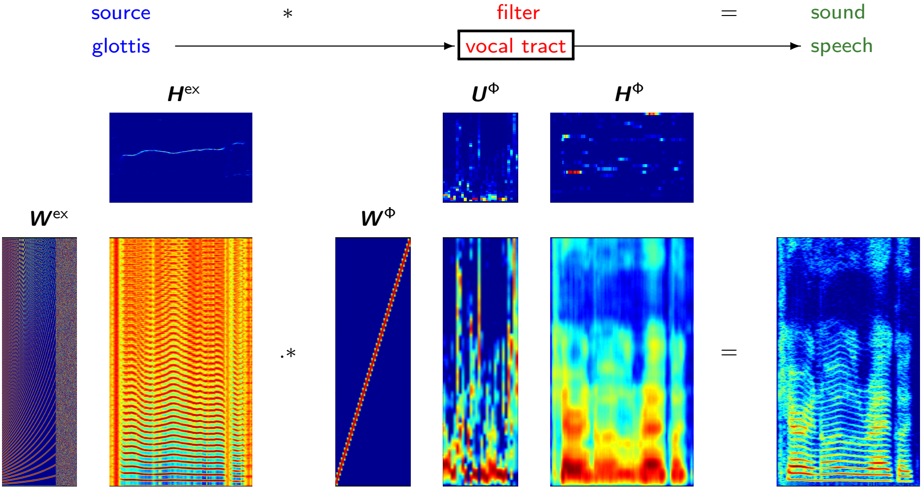


=



$$\left(\boldsymbol{W}^{\text{ex}} \quad \boldsymbol{H}^{\text{ex}}\right) \quad .* \quad \left(\widehat{\boldsymbol{W}}^{\Phi} \quad \boldsymbol{H}^{\Phi}\right) \quad = \quad \boldsymbol{V}$$

Source/filter model [Durrieu et al., 2009]

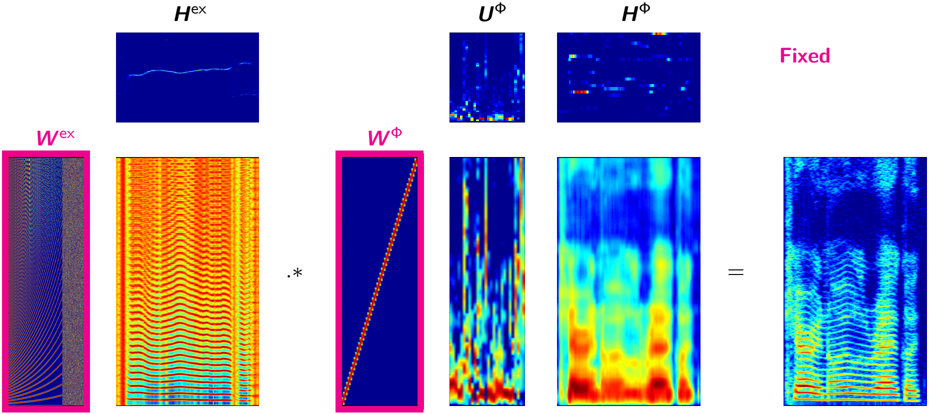


$$(\boldsymbol{W}^{\mathrm{ex}} \quad \boldsymbol{H}^{\mathrm{ex}}) \quad .* \quad \left(\boldsymbol{W}^{\Phi} \quad \boldsymbol{U}^{\Phi} \quad \boldsymbol{H}^{\Phi}\right) \quad = \quad \boldsymbol{V}$$

Source/filter model [Durrieu et al., 2009]

source * filter = sound

glottis → vocal tract → speech

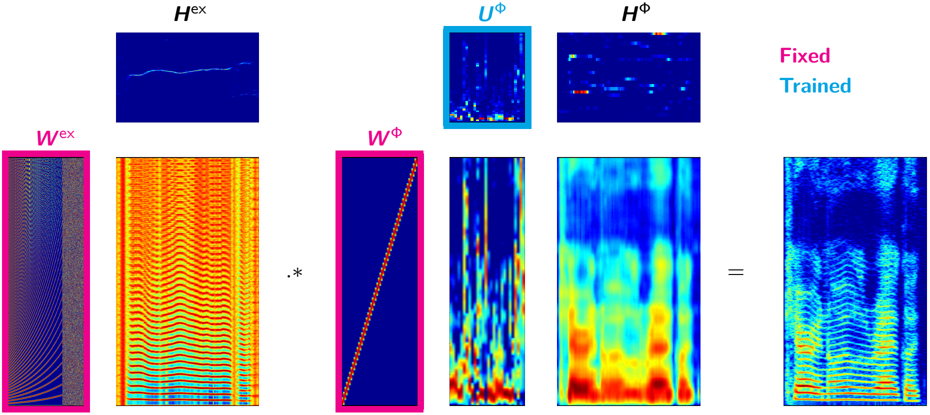


$$(\boldsymbol{W}^{\mathrm{ex}} \boldsymbol{H}^{\mathrm{ex}}) .* (\boldsymbol{W}^{\Phi} \boldsymbol{U}^{\Phi} \boldsymbol{H}^{\Phi}) = \boldsymbol{V}$$

Source/filter model [Durrieu et al., 2009]

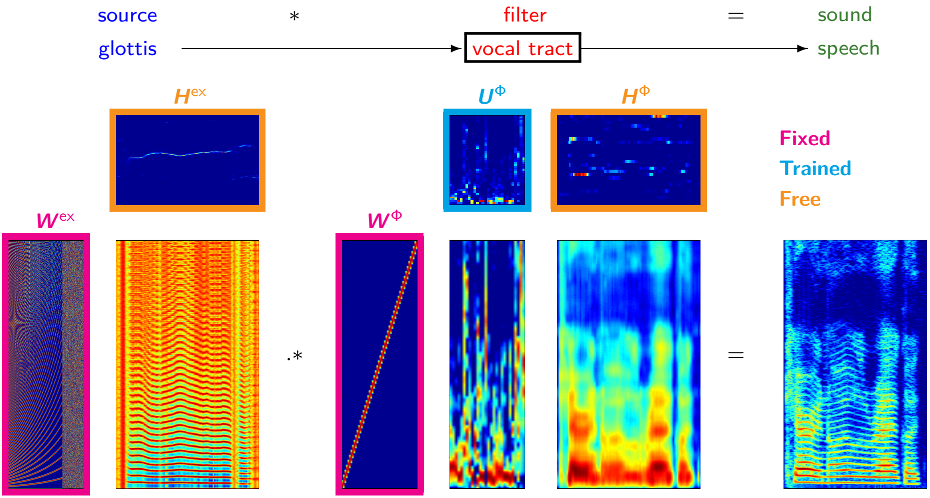
source * filter = sound

glottis → vocal tract → speech



$$(W^{ex} \quad H^{ex}) \; .* \; (W^{\Phi} \quad U^{\Phi} \quad H^{\Phi}) = V$$

Source/filter model [Durrieu et al., 2009]



$$(W^{ex} \quad H^{ex}) \;.*\; (W^{\Phi} \quad U^{\Phi} \quad H^{\Phi}) = V$$

State of the art

- Speech separation using NMF
- Semi-supervised NMF
- Source/filter model

Proposed method

- Semi-supervised constrained NMF
- Contribution 1: speech-specific source/filter coherence constraint
- Contribution 2: adaptive weight method

Experimental evaluation

- Experiment description
- Effect of weight's adaptation
- Algorithm comparison

Semi-supervised constrained NMF

Physically-informed model

$$\underset{\boldsymbol{H}^{\text{ex}},\boldsymbol{H}^{\Phi},\boldsymbol{W}^{N},\boldsymbol{H}^{N}\geq 0}{\operatorname{argmin}} \mathcal{C}(\boldsymbol{V}|\widetilde{\boldsymbol{V}}) \text{ with } \begin{cases} \widetilde{\boldsymbol{V}} = \boldsymbol{W}^{\text{ex}}\boldsymbol{H}^{\text{ex}} \otimes \boldsymbol{W}^{\Phi}\boldsymbol{U}^{\Phi}\boldsymbol{H}^{\Phi} + \boldsymbol{W}^{N}\boldsymbol{H}^{N} \\ \boldsymbol{W}^{\text{ex}} \text{ and } \boldsymbol{W}^{\Phi} \text{ fixed} \\ \boldsymbol{U}^{\Phi} \text{ learned} \end{cases} \quad (2)$$

But still no physically-coherent behavior.

Semi-supervised constrained NMF

Physically-informed model

$$\underset{\boldsymbol{H}^{\text{ex}},\boldsymbol{H}^{\Phi},\boldsymbol{W}^{N},\boldsymbol{H}^{N}\geq 0}{\text{argmin}} \; \mathcal{C}(\boldsymbol{V}|\widetilde{\boldsymbol{V}}) \text{ with } \begin{cases} \widetilde{\boldsymbol{V}} = \boldsymbol{W}^{\text{ex}}\boldsymbol{H}^{\text{ex}} \otimes \boldsymbol{W}^{\Phi}\boldsymbol{U}^{\Phi}\boldsymbol{H}^{\Phi} + \boldsymbol{W}^{N}\boldsymbol{H}^{N} \\ \boldsymbol{W}^{\text{ex}} \text{ and } \boldsymbol{W}^{\Phi} \text{ fixed} \\ \boldsymbol{U}^{\Phi} \text{ learned} \end{cases} \tag{2}$$

But still no physically-coherent behavior.

Constraints for controlling its behavior

$$\underbrace{\mathcal{C}(\boldsymbol{V}|\widetilde{\boldsymbol{V}})}_{\text{Total cost}} = \underbrace{D(\boldsymbol{V}|\widetilde{\boldsymbol{V}})}_{\text{Reconstruction cost}} + \overbrace{\lambda}^{\text{Weight parameter}} \underbrace{\mathcal{P}(\boldsymbol{\Theta})}_{\text{Constraint penalty}} \tag{3}$$

Semi-supervised constrained NMF

Physically-informed model

$$\underset{\boldsymbol{H}^{\text{ex}},\boldsymbol{H}^{\Phi},\boldsymbol{W}^{N},\boldsymbol{H}^{N}\geq 0}{\text{argmin}} \ \mathcal{C}(\boldsymbol{V}|\widetilde{\boldsymbol{V}}) \text{ with } \begin{cases} \widetilde{\boldsymbol{V}} = \boldsymbol{W}^{\text{ex}}\boldsymbol{H}^{\text{ex}} \otimes \boldsymbol{W}^{\Phi}\boldsymbol{U}^{\Phi}\boldsymbol{H}^{\Phi} + \boldsymbol{W}^{N}\boldsymbol{H}^{N} \\ \boldsymbol{W}^{\text{ex}} \text{ and } \boldsymbol{W}^{\Phi} \text{ fixed} \\ \boldsymbol{U}^{\Phi} \text{ learned} \end{cases} \tag{2}$$

But still no physically-coherent behavior.

Constraints for controlling its behavior

$$\underbrace{\mathcal{C}(\boldsymbol{V}|\widetilde{\boldsymbol{V}})}_{\text{Total cost}} = \underbrace{D(\boldsymbol{V}|\widetilde{\boldsymbol{V}})}_{\text{Reconstruction cost}} + \overbrace{\lambda}^{\text{Weight parameter}} \underbrace{\mathcal{P}(\boldsymbol{\Theta})}_{\text{Constraint penalty}} \tag{3}$$

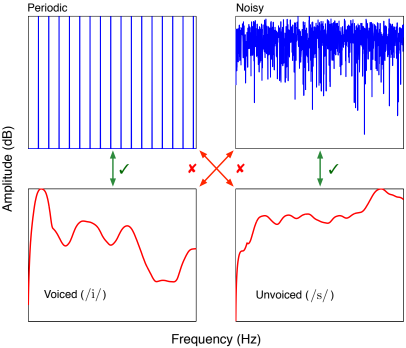
New multiplicative update rules :

$$\boldsymbol{\Theta}^{(i+1)} \longleftarrow \boldsymbol{\Theta}^{(i)} \otimes \frac{\nabla^{-}_{\boldsymbol{\Theta}} D + \lambda \nabla^{-}_{\boldsymbol{\Theta}} \mathcal{P}}{\nabla^{+}_{\boldsymbol{\Theta}} D + \lambda \nabla^{+}_{\boldsymbol{\Theta}} \mathcal{P}} \quad \forall \boldsymbol{\Theta} \in \left\{ \boldsymbol{H}^{\text{ex}}, \boldsymbol{H}^{\Phi}, \boldsymbol{W}^{N}, \boldsymbol{H}^{N} \right\} \tag{4}$$

Constraints from literature [Bertin, 2009]: Sparsity, Decorrelation, Smoothness.

Contribution 1: speech-specific source/filter coherence constraint

Problem: unrealistic source/filter combinations possible



We only want to allow:

- periodic excitation with adequate filter (e.g. vowels, voice consonants)
- noisy excitation with adequate filter (e.g., unvoiced consonants)

Contribution 1: speech-specific source/filter coherence constraint

New constraint (that requires phoneme-labelled spectral filter basis)

$$\mathcal{P}_{\phi}(\boldsymbol{H}^{\text{ex}}, \boldsymbol{H}^{\Phi}) \tag{5}$$

Contribution 1: speech-specific source/filter coherence constraint

New constraint (that requires phoneme-labelled spectral filter basis)

$$\mathcal{P}_\phi(\boldsymbol{H}^{\mathrm{ex}}, \boldsymbol{H}^{\Phi}) = \sum_{\substack{k \in \mathrm{periodics} \\ l \in \mathrm{unvoiced}}} \frac{\left[\boldsymbol{H}^{\mathrm{ex}} \boldsymbol{H}^{\Phi\,T}\right]_{kl}}{} \quad (5)$$

- : measure of correlation

Contribution 1: speech-specific source/filter coherence constraint

New constraint (that requires phoneme-labelled spectral filter basis)

$$\mathcal{P}_{\phi}(\boldsymbol{H}^{\text{ex}}, \boldsymbol{H}^{\Phi}) = \sum_{\substack{k \in \text{periodics} \\ l \in \text{unvoiced}}} \frac{\left[\boldsymbol{H}^{\text{ex}} \boldsymbol{H}^{\Phi\,T} \right]_{kl}}{\|\boldsymbol{H}^{\text{ex}}_k\|_{\ell_2} \|\boldsymbol{H}^{\Phi}_l\|_{\ell_2}} \tag{5}$$

- : measure of correlation
- : normalized

Contribution 1: speech-specific source/filter coherence constraint

New constraint (that requires phoneme-labelled spectral filter basis)

$$\mathcal{P}_\phi(\boldsymbol{H}^{\text{ex}}, \boldsymbol{H}^{\Phi}) = \sum_{\substack{k \in \text{periodics} \\ l \in \text{unvoiced}}} \frac{\left[\boldsymbol{H}^{\text{ex}} \boldsymbol{H}^{\Phi\,T}\right]_{kl}}{\|\boldsymbol{H}_k^{\text{ex}}\|_{\ell_2} \|\boldsymbol{H}_l^{\Phi}\|_{\ell_2}} + \sum_{\substack{k \in \text{noisy} \\ l \in \text{voiced}}} \frac{\left[\boldsymbol{H}^{\text{ex}} \boldsymbol{H}^{\Phi\,T}\right]_{kl}}{\|\boldsymbol{H}_k^{\text{ex}}\|_{\ell_2} \|\boldsymbol{H}_l^{\Phi}\|_{\ell_2}} \tag{5}$$

- : measure of correlation
- : normalized
- : for both type of unwanted combination

Contribution 2: adaptive weight method

Main issue with constrained NMF

Adjusting the weight parameter λ :

- if too small, no effect is visible;
- if too big, convergence becomes extremely sensitive to initialization (which is typically random).

Contribution 2: adaptive weight method

Main issue with constrained NMF

Adjusting the weight parameter λ :

- if too small, no effect is visible;
- if too big, convergence becomes extremely sensitive to initialization (which is typically random).

Idea

Adjust the constraint weight at each iteration of the NMF:

- constraint relaxed during strong evolution of the reconstruction cost;
- constraint enforced when the reconstruction is more stable;

Contribution 2: adaptive weight method

Main issue with constrained NMF

Adjusting the weight parameter λ :

- if too small, no effect is visible;
- if too big, convergence becomes extremely sensitive to initialization (which is typically random).

Idea

Adjust the constraint weight at each iteration of the NMF:

- constraint relaxed during strong evolution of the reconstruction cost;
- constraint enforced when the reconstruction is more stable;

Adaptive method

$$\lambda^{(i)} = \lambda_{Max} \frac{D(\boldsymbol{V}|\widetilde{\boldsymbol{V}}^{(i-1)})}{D(\boldsymbol{V}|\widetilde{\boldsymbol{V}}^{(i-2)})} \tag{6}$$

$$\left(\begin{array}{c} D(\boldsymbol{V}|\widetilde{\boldsymbol{V}}) \text{ the reconstruction cost} \\ \lambda \in [0 \ \lambda_{Max}] \end{array} \right)$$

State of the art

- Speech separation using NMF
- Semi-supervised NMF
- Source/filter model

Proposed method

- Semi-supervised constrained NMF
- Contribution 1: speech-specific source/filter coherence constraint
- Contribution 2: adaptive weight method

Experimental evaluation

- Experiment description
- Effect of weight's adaptation
- Algorithm comparison

Experiment description

Database

TIMIT [Zue et al., 1990] : 20 speakers, 2 learnings sentences and 8 test sentences
QUT-NOISE [Dean et al., 2010] : 5 types of noise
Mixed at 3 Signal-to-Noise Ratio ($-6\mathrm{dB}$, $+0\mathrm{dB}$ and $+6\mathrm{dB}$)

Benchmark

SoA	• ASNA [Virtanen et al., 2013]	supervised
	• IMM [Durrieu et al., 2009]	unsupervised
Proposed	• S-IMM: without contrainsts	semi-supervised
	• SC-IMM1: state-of-the art contrainsts	
	• SC-IMM2: source/filter coherence constraint	
	• SC-IMM3: all constraints	

Measures

- SDR : Signal to Distortion Ratio (in dB)
- PESQ : Perceptual Evaluation of Speech Quality (from 1 (bad) to 5 (excellent))

Effect of weight's adaptation

SNR	Measure	With adaptation			Without adaptation		
		SC-IMM1	SC-IMM2	SC-IMM3	SC-IMM1	SC-IMM2	SC-IMM3
-6dB	SDR (dB)	4.1	**5.2**	**5.4**	4.1	5.0	5.2
	PESQ	1.91	**2.01**	**2.01**	1.91	1.94	1.92
$+0$dB	SDR (dB)	9.2	**9.8**	**9.8**	9.2	9.0	8.9
	PESQ	2.30	**2.34**	**2.35**	2.30	2.24	2.23
$+6$dB	SDR (dB)	**13.0**	**12.8**	**12.9**	12.8	11.1	10.9
	PESQ	**2.62**	**2.59**	**2.62**	2.61	2.46	2.44
Mean	SDR (dB)	8.7	**9.3**	**9.4**	8.7	8.4	8.3
	PESQ	**2.28**	**2.31**	**2.33**	2.27	2.21	2.20

$\Rightarrow$ **adaptation gives best results**

Algorithm comparison

SNR	Measure	Algorithms					
		ASNA	IMM	S-IMM	SC-IMM1	SC-IMM2	SC-IMM3
$-6\mathrm{dB}$	SDR (dB)	5.8	4.4	4.0	4.1	5.2	5.4
	PESQ	2.00	1.22	1.91	1.91	2.01	2.01
$+0\mathrm{dB}$	SDR (dB)	10.7	7.8	9.1	9.2	9.8	9.8
	PESQ	2.44	1.54	2.30	2.30	2.34	2.35
$+6\mathrm{dB}$	SDR (dB)	15.0	9.7	13.0	13.0	12.8	12.9
	PESQ	2.85	1.82	2.62	2.62	2.59	2.62
Mean	SDR (dB)	10.5	7.3	8.7	8.7	9.3	9.4
	PESQ	2.43	1.52	2.28	2.28	2.31	2.33

Algorithm comparison

SNR	Measure	Algorithms					
		ASNA	IMM	S-IMM	SC-IMM1	SC-IMM2	SC-IMM3
-6dB	SDR (dB)	5.8	4.4	4.0	4.1	5.2	5.4
	PESQ	2.00	1.22	**1.91**	1.91	2.01	2.01
$+0$dB	SDR (dB)	10.7	7.8	**9.1**	9.2	9.8	9.8
	PESQ	2.44	1.54	**2.30**	2.30	2.34	2.35
$+6$dB	SDR (dB)	15.0	9.7	**13.0**	13.0	12.8	12.9
	PESQ	2.85	1.82	**2.62**	2.62	2.59	2.62
Mean	SDR (dB)	10.5	7.3	**8.7**	8.7	9.3	9.4
	PESQ	2.43	1.52	**2.28**	2.28	2.31	2.33

$\Rightarrow$ **supervisation helps separation**

Algorithm comparison

SNR	Measure	ASNA	IMM	S-IMM	SC-IMM1	SC-IMM2	SC-IMM3
-6dB	SDR (dB)	5.8	4.4	4.0	4.1	5.2	**5.4**
	PESQ	2.00	1.22	1.91	1.91	**2.01**	**2.01**
$+0$dB	SDR (dB)	10.7	7.8	9.1	9.2	**9.8**	**9.8**
	PESQ	2.44	1.54	2.30	2.30	2.34	**2.35**
$+6$dB	SDR (dB)	15.0	9.7	**13.0**	**13.0**	12.8	12.9
	PESQ	2.85	1.82	**2.62**	**2.62**	2.59	**2.62**
Mean	SDR (dB)	10.5	7.3	8.7	8.7	9.3	**9.4**
	PESQ	2.43	1.52	2.28	2.28	2.31	**2.33**

$\Rightarrow$ **best proposed algorithm: SC-IMM3 (with all constraints)**

Algorithm comparison

SNR	Measure	ASNA	IMM	S-IMM	SC-IMM1	SC-IMM2	SC-IMM3
-6dB	SDR (dB)	5.8	4.4	4.0	4.1	5.2	5.4
	PESQ	2.00	1.22	1.91	1.91	2.01	2.01
$+0$dB	SDR (dB)	10.7	7.8	9.1	9.2	9.8	9.8
	PESQ	2.44	1.54	2.30	2.30	2.34	2.35
$+6$dB	SDR (dB)	15.0	9.7	13.0	13.0	12.8	12.9
	PESQ	2.85	1.82	2.62	2.62	2.59	2.62
Mean	SDR (dB)	10.5	7.3	8.7	8.7	9.3	9.4
	PESQ	2.43	1.52	2.28	2.28	2.31	2.33

$\Rightarrow$ **better than Durrieu & close of Virtanen in low SNRs**

Speech

Noise

Text : *"Computers are being used to keep branch inventories at more workable levels."*

Conclusion

Summary

- Semi-supervised speech separation
- Source/filter model

Contributions

- Weight adaptation method for constraints
- Source/filter coherence constraint for speech
- Good results close to literature in supervised separation

Further research

- Speaker-independant model [Sun and Mysore, 2013]
- Integration of a language model [Mysore and Smaragdis, 2012]
- Integration of a noise adaptation method [Roebel et al., 2015]

Bibliography I

Gautham J. Mysore and Paris Smaragdis. A non-negative approach to language informed speech separation. In *Latent Variable Analysis and Signal Separation*, pages 356–363. Springer, 2012.

Tuomas Virtanen, Jort F Gemmeke, and Bhiksha Raj. Active-set Newton algorithm for overcomplete non-negative representations of audio. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2277–2289, 2013.

Dennis L. Sun and Gautham J. Mysore. Universal speech models for speaker independent single channel source separation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 141–145, 2013.

François G. Germain and Gautham J. Mysore. Speaker and noise independent online single-channel speech enhancement. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 71–75, 2015.

Luc Le Magoarou, Alexey Ozerov, and Ngoc QK Duong. Text-informed audio source separation. Example-based approach using non-negative matrix partial co-factorization. *Journal of Signal Processing Systems*, pages 1–15, 2014.

Jean-Louis Durrieu, Alexey Ozerov, Cédric Févotte, Gaël Richard, and Bertrand David. Main instrument separation from stereophonic audio signals using a source/filter model. In *European Signal Processing Conference (EUSIPCO)*, pages 15–19, 2009. URL `http://www.durrieu.ch/phd/eusipco09/`.

Bibliography II

Nancy Bertin. *Les factorisations en matrices non-négatives. Approches contraintes et probabilistes, application à la transcription automatique de musique polyphonique.* PhD thesis, Télécom ParisTech, 2009.

Victor Zue, Stephanie Seneff, and James Glass. Speech database development at MIT: TIMIT and beyond. *Speech Communication*, 9(4):351–356, 1990.

David B. Dean, Sridha Sridharan, Robert J. Vogt, and Michael W. Mason. The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms. *Proceedings of Interspeech 2010*, pages 3110–3113, 2010.

Axel Roebel, Jordi Pons, Marco Liuni, and Mathieu Lagrange. On automatic drum transcription using non-negative matrix deconvolution and Itakura-Saito divergence. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 414–418, 2015.