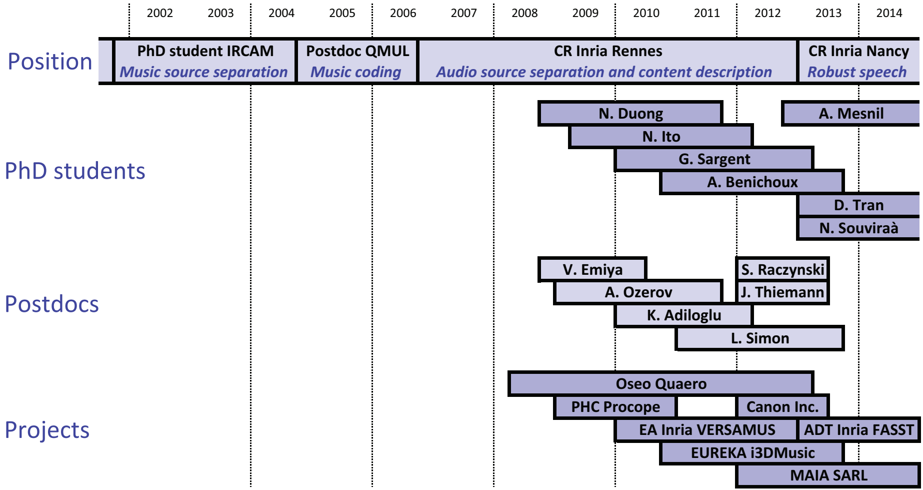
Contributions to audio source separation and content description

E. Vincent

METISS Team, Inria Rennes - Bretagne Atlantique

Career path



2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
Position
PhD student IRCAM
Music source separation
Postdoc QMUL
Music coding
CR Inria Rennes
Audio source separation and content description
CR Inria Nancy
Robust speech
PhD students
N. Duong
A. Mesnil
N. Ito
G. Sargent
A. Benichoux
D. Tran
N. Souviraà
Postdocs
V. Emiya
S. Raczynski
A. Ozerov
J. Thiemann
K. Adiloglu
L. Simon
Projects
Oseo Quaero
PHC Procope
Canon Inc.
EA Inria VERSAMUS
ADT Inria FASST
EUREKA i3DMusic
MAIA SARL

Audio in the real world

The audio modality is essential in daily situations: spoken communication, TV, music, entertainment. . .

But audio scenes are often more complex than we would like!

Ex: TV series

Audio in the real world

The audio modality is essential in daily situations: spoken communication, TV, music, entertainment...

But audio scenes are often more complex than we would like!

Ex: TV series

Many sound sources: Speech, music, background noise.

Audio in the real world

The audio modality is essential in daily situations: spoken communication, TV, music, entertainment. . .

But audio scenes are often more complex than we would like!

Ex: TV series

Many sound sources: Speech, music, background noise.

Much information: Who is speaking? What is he saying?

Audio in the real world

The audio modality is essential in daily situations: spoken communication, TV, music, entertainment...

But audio scenes are often more complex than we would like!

Ex: TV series

Many sound sources: Speech, music, background noise.

Much information: Who is speaking? What is he saying?

Where is he? How stressed is he?

Audio in the real world

The audio modality is essential in daily situations: spoken communication, TV, music, entertainment...

But audio scenes are often more complex than we would like!

Ex: TV series

Many sound sources: Speech, music, background noise.

Much information: Who is speaking? What is he saying?

Where is he? How stressed is he?

What's the music style? The bombing rate?

Audio in the real world

The audio modality is essential in daily situations: spoken communication, TV, music, entertainment. . .

But audio scenes are often more complex than we would like!

Ex: TV series

Many sound sources: Speech, music, background noise.

Much information: Who is speaking? What is he saying?

Where is he? How stressed is he?

What's the music style? The bombing rate?

What is happening? What's gonna happen next? 

General goal and stakes

We want to:

- enhance the sound sources of interest
- extract the corresponding information

Wide range of applications, including:

- high-fidelity hearing aids and mobile communications,
- voice applications, multimedia document indexing, music search,
- 3D audio rendering, repurposing, interactive applications. . .

General goal and stakes

We want to:

- enhance the sound sources of interest: source separation
- extract the corresponding information: content description

Wide range of applications, including:

- high-fidelity hearing aids and mobile communications,
- voice applications, multimedia document indexing, music search,
- 3D audio rendering, repurposing, interactive applications. . .

Part 1. Audio source separation

Part 2. Audio content description

Part 3. Research directions

Part 1. Audio source separation

Part 2. Audio content description

Part 3. Research directions

Audio source separation: the basics

Additive mixing:

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t)$$

$\mathbf{x}(t)$: multichannel mixture
$\mathbf{c}_j(t)$: jth spatial source image

(Not so) special case: point sources

$$\mathbf{c}_j(t) = \mathbf{a}_j \star s_j(t)$$

$\mathbf{a}_j(\tau)$: mixing filter
$s_j(t)$: jth source signal

Goal: estimate $\mathbf{c}_j(t)$ given $\mathbf{x}(t)$.

Evolution of the research focus

2012

Real-world
unknown nb
of sources

Real-world
e.g., diffuse

Convolutive
echoic

Under-determined
nb sources >
mixture channels

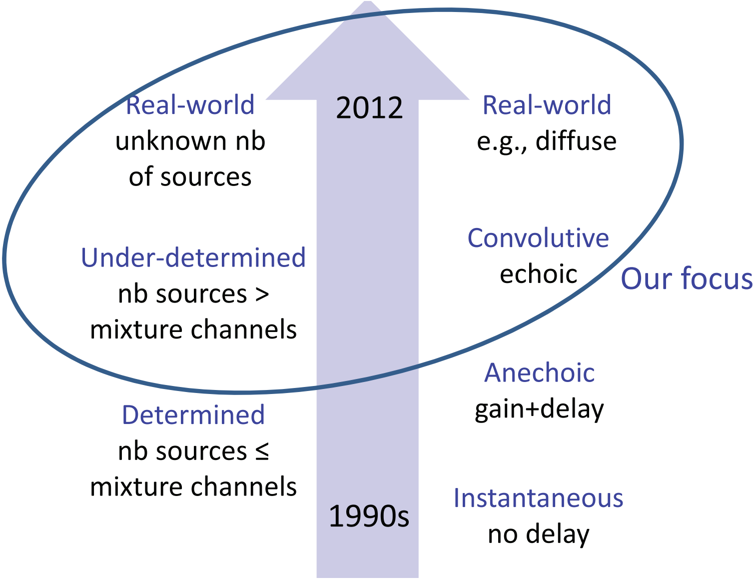
Anechoic
gain+delay

Determined
nb sources ≤
mixture channels

Instantaneous
no delay

1990s

Evolution of the research focus



Spatial and spectral cues (1)

Standard principle:

- work in the time-frequency domain

$$\widetilde{\mathbf{x}}(n,f) = \sum_{j=1}^{J} \widetilde{\mathbf{c}}_j(n,f)$$

$\widetilde{\mathbf{x}}(n,f)$: vector of mixture TF coeff.
$\widetilde{\mathbf{c}}_j(n,f)$: jth source spatial image TF coeff.

- for point sources, replace convolution by narrowband multiplication

$$\widetilde{\mathbf{c}}_j(n,f) = \widetilde{\mathbf{a}}_j(f)\,\widetilde{s}_j(n,f)$$

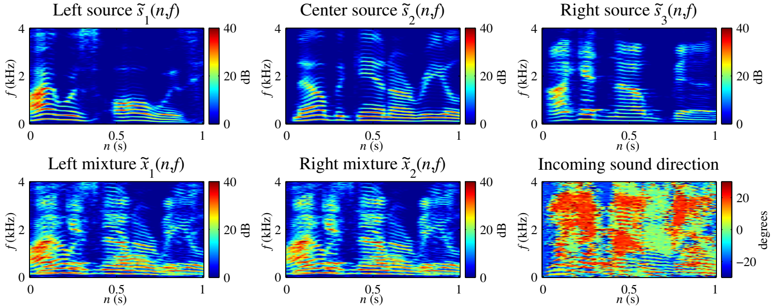
$\widetilde{\mathbf{c}}_j(n,f)$: jth source spatial image TF coeff.
$\widetilde{\mathbf{a}}_j(f)$: mixing coefficients
$\widetilde{s}_j(n,f)$: jth source TF coeff.

- ...

Spatial and spectral cues (2)

Standard principle:

- ...
- estimate $\widetilde{\mathbf{a}}_j(f)$ and $\widetilde{s}_j(n,f)$ by time-frequency clustering of spatial cues [Zibulevsky, Rickard, Gribonval...]



- exploit additional spectral cues to separate overlapping sources or sources from the same direction [Benaroya, Virtanen, Vincent...]

Contributions and positioning

TF representation	STFT		Adaptive [Vincent et al., 2005]	Auditory-motivated [PhD Duong]
Spatial model	Instantaneous	Anechoic	Convolutive	Full-rank [PhD Duong]
			Wideband [Kowalski et al., 2010]	
Spectral model	Binary	Gaussian	GMM	NMF
	ℓ_1	ℓ_p [Vincent et al., 2007]		Multilevel NMF [postdoc Ozerov]
Estimation	Two-stage	Joint ML	Convex penalties [PhD Ito, Benichoux]	MAP [PhD Duong]
⋮	Online EM [postdoc L. Simon]		Fast EM [postdoc J. Thiemann]	Consistent Wiener [NTT patent]

Contributions and positioning

TF representation	STFT		Adaptive [Vincent et al., 2005]	Auditory-motivated [PhD Duong]
Spatial model	Instantaneous	Anechoic	Convolutive	Full-rank [PhD Duong]
				Wideband [Kowalski et al., 2010]
Spectral model	Binary, ℓ_1	Gaussian, ℓ_p [Vincent et al., 2007]	GMM	NMF, Multilevel NMF [postdoc Ozerov]
Estimation	Two-stage	Joint ML	Convex penalties [PhD Ito, Benichoux]	MAP [PhD Duong]
⋮	Online EM [postdoc L. Simon]		Fast EM [postdoc J. Thiemann]	Consistent Wiener [NTT patent]

The rank-1 spatial model

Former state-of-the-art: narrowband approximation

$$\widetilde{\mathbf{c}}_j(n,f) = \widetilde{\mathbf{a}}_j(f)\,\widetilde{s}_j(n,f)$$

$\widetilde{\mathbf{c}}_j(n,f)$: jth source spatial image TF coeff.
$\widetilde{\mathbf{a}}_j(f)$: Fourier transform of $\mathbf{a}_j(\tau)$
$\widetilde{s}_j(n,f)$: jth source TF coeff.

In the Gaussian (variance) modeling framework,

$$\widetilde{s}_j(n,f) \sim \mathcal{N}(0, v_j(n,f)) \quad \Rightarrow \quad \mathbf{c}_j(n,f) \sim \mathcal{N}(0, v_j(n,f)\widetilde{\mathbf{a}}_j(f)\widetilde{\mathbf{a}}_j(f)^H)$$

This rank-1 model essentially represents the apparent spatial direction of sound at frequency f.

Problem: reverberation induces echoes from all directions. The notion of mixing filter $\mathbf{a}_j(\tau)$ does not even make sense for diffuse sources.

Proposed full-rank spatial model

Proposed model [PhD Duong]:

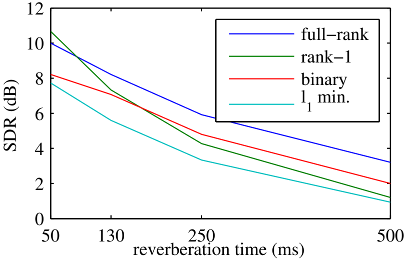
$$\mathbf{c}_j(n,f) \sim \mathcal{N}(0, v_j(n,f)\mathbf{\Sigma}_j(f))$$

with $\mathbf{\Sigma}_j(f)$ full-rank spatial covariance matrix.

Represents both the spatial direction and the spatial width of the source.

Derived an expectation-maximization (EM) algorithm for ML estimation.

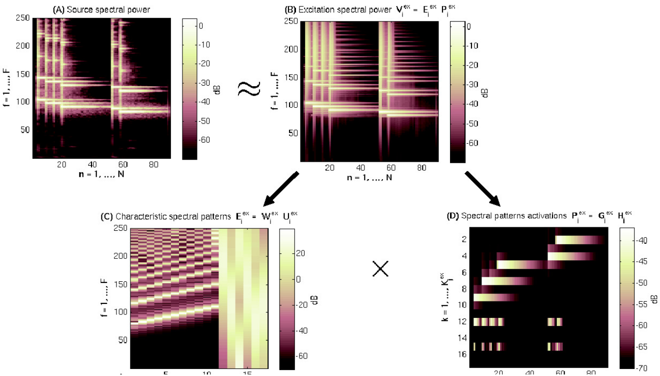
Results on two-channel mixtures of three sources



Conventional NMF

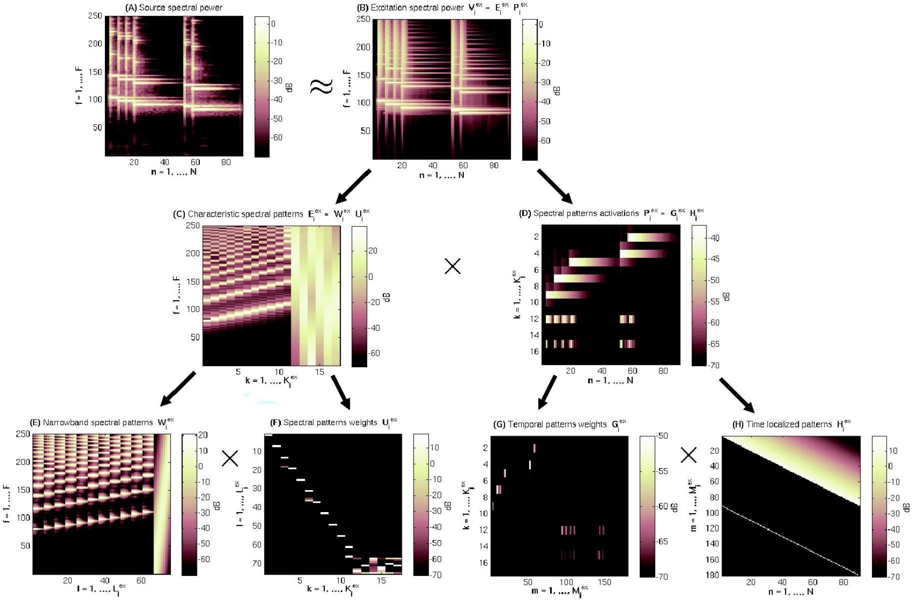
Former state-of-the-art: nonnegative matrix factorization (NMF)

$$v_j(n, f) = \sum_k w_{jk}(f) h_{jk}(n)$$



Problem: either too rigid ($w_{jk}(f)$ fixed) or prone to overfitting ($w_{jk}(f)$ adaptive).

Proposed multilevel NMF (1) [Vincent 2007, postdoc Ozerov]



Proposed multilevel NMF (2)

Can handle new constraints: harmonicity, smooth envelope, attack type...

Derived a flexible EM algorithm for joint estimation of all layers, whether fixed or adaptive ⇒ **FASST Toolbox**.

Results on two-channel mixtures of three or four sources (SiSEC 2010)

Spatial, spectral, and temporal constraints			Average SDR (dB)	
rank	spec	temp	5 cm	1 m
1			2.2	2.5
2			2.0	3.0
1	X		2.2	2.8
2	X		2.3	3.2
1		X	2.4	2.6
2		X	2.1	2.9
1	X	X	**2.5**	3.9
2	X	X	2.3	**5.0**

Also best general algorithm for the separation of music recordings in SiSEC 2011.

Evaluation: a transversal activity

Complete evaluation methodology for audio source separation:

- formalization of audio source separation tasks [Vincent et al., 2007]
- definition of objective/subjective evaluation criteria [postdoc Emiya]
  $\Rightarrow$ **BSS Eval & PEASS Toolboxes**
- computation of theoretical performance bounds [Vincent et al., 2007].

Co-founded two series of evaluation campaigns:

- SASSEC/SiSEC (source separation): 119 entries since 2007
- CHiME (noise-robust speech recognition): 13 in 2011, again in 2013

Impact:

- helped the adoption of common problems, datasets and metrics,
- helped focus on the remaining challenges: lack of spatial diversity, reverberation, source movements, background noise.

Are we there yet?

mix   vocals   drums   bass   piano
(separation by FASST)

Are we there yet?

mix   vocals   drums   bass   piano
(separation by FASST)

This level of quality is sufficient for many signal enhancement/remixing applications.

Ongoing industrial transfer to Canon Inc., Audionamix SA and MAIA SARL.

Are we there yet?

mix vocals drums bass piano
(separation by FASST)

This level of quality is sufficient for many signal enhancement/remixing applications.

Ongoing industrial transfer to Canon Inc., Audionamix SA and MAIA SARL.

Is it sufficient for content description?

Part 1. Audio source separation

Part 2. Audio content description

Part 3. Research directions

Audio content description: the basics

Audio content description techniques do not operate on the signals directly but on derived features, e.g., Mel frequency cepstral coefficients (MFCCs).

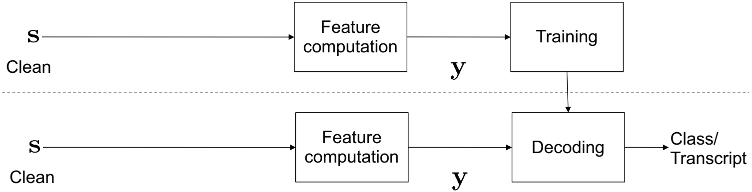
Classification/transcription most often relies on probabilistic acoustic models of the features, e.g., Gaussian mixture models (GMMs).

Audio content description: the basics

Audio content description techniques do not operate on the signals directly but on derived features, e.g., Mel frequency cepstral coefficients (MFCCs).

Classification/transcription most often relies on probabilistic acoustic models of the features, e.g., Gaussian mixture models (GMMs).

Two stages: training and decoding.

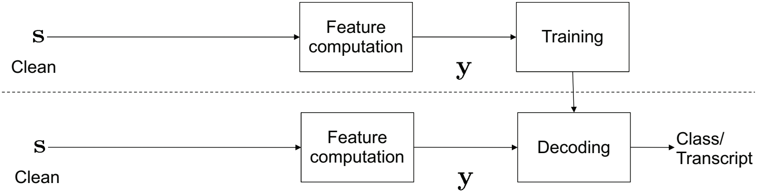


Audio content description: the basics

Audio content description techniques do not operate on the signals directly but on derived features, e.g., Mel frequency cepstral coefficients (MFCCs).

Classification/transcription most often relies on probabilistic acoustic models of the features, e.g., Gaussian mixture models (GMMs).

Two stages: training and decoding.

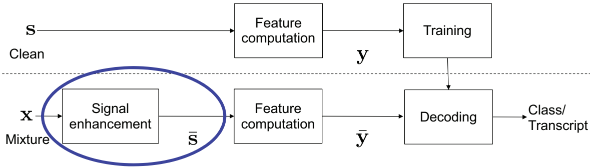


Problem: matched training/test paradigm, works only for clean data.

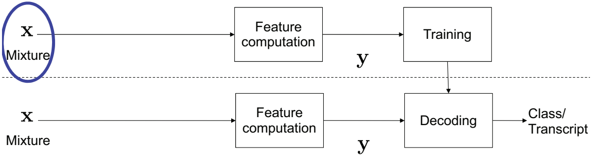
How can we reduce the mismatch for noisy/mixture data?

Conventional techniques for noise robustness

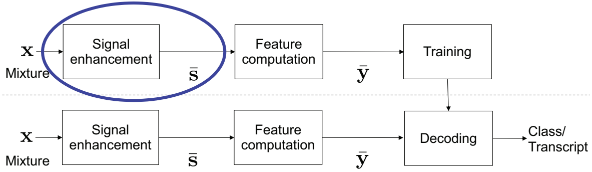
Feature compensation: good separation but often increased mismatch



Training data coverage: better match but huge training set needed

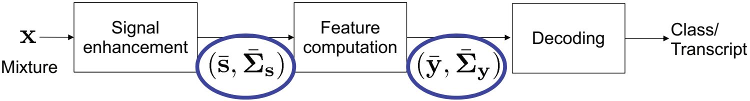


Noise adaptive training [Deng, 2000]: combines both advantages, large training set still needed



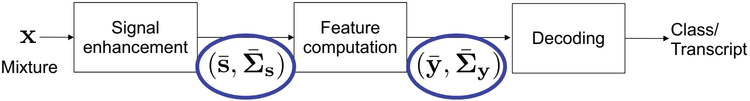
Uncertainty propagation and decoding

Emerging paradigm: estimate and propagate confidence values represented by Gaussian posterior distributions [Deng, Astudillo, Kolossa. . .].



Uncertainty propagation and decoding

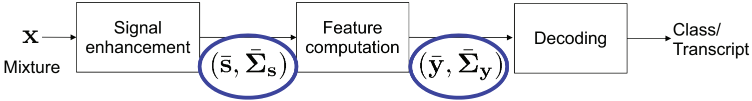
Emerging paradigm: estimate and propagate confidence values represented by Gaussian posterior distributions [Deng, Astudillo, Kolossa. . .].



$\bar{\boldsymbol{\Sigma}}_{\mathbf{s}}$	ML or heuristic	Bayesian [postdoc Adiloğlu]
$(\mathbf{s}, \bar{\boldsymbol{\Sigma}}_{\mathbf{s}}) \rightarrow (\mathbf{y}, \bar{\boldsymbol{\Sigma}}_{\mathbf{y}})$	Moment matching, unscented transform. . .	
Decoding	Uncertainty decoding, modified imputation. . .	
Training	Clean data	Uncertainty training [postdoc Ozerov]

Uncertainty propagation and decoding

Emerging paradigm: estimate and propagate confidence values represented by Gaussian posterior distributions [Deng, Astudillo, Kolossa...].



$\bar{\boldsymbol{\Sigma}}_{\mathbf{s}}$	ML or heuristic	Bayesian [postdoc Adiloğlu]
$(\mathbf{s}, \bar{\boldsymbol{\Sigma}}_{\mathbf{s}}) \rightarrow (\mathbf{y}, \bar{\boldsymbol{\Sigma}}_{\mathbf{y}})$	Moment matching, unscented transform...	
Decoding	Uncertainty decoding, modified imputation...	
Training	Clean data	Uncertainty training [postdoc Ozerov]

Bayesian uncertainty estimator (1)

Conventional ML uncertainty estimator:

$$p(\mathbf{s}|\mathbf{x}) = p(\mathbf{s}|\mathbf{x}, \widehat{\theta}) \quad \text{with} \quad \widehat{\theta} = \arg\max p(\mathbf{x}|\theta)$$

Proposed Bayesian uncertainty estimator [postdoc Adiloğlu]:

$$p(\mathbf{s}|\mathbf{x}) = \int p(\mathbf{s}, \theta|\mathbf{x})\, d\theta$$

$\mathbf{s}$: target source STFT coeff.
$\mathbf{x}(t)$: mixture STFT coeff.
θ: separation model parameters

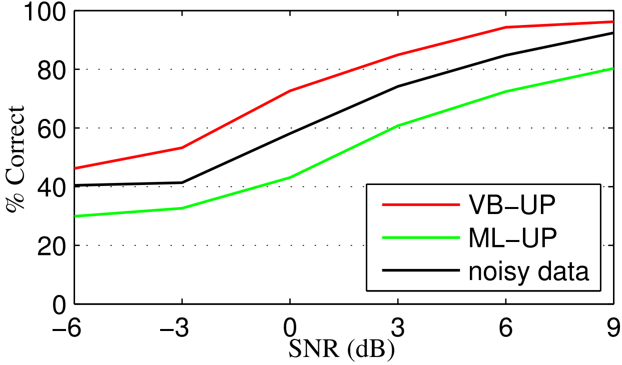
Derived a tractable variational Bayesian (VB) EM approximation.

Similar to conventional ML-EM, but update posterior parameter distributions instead of parameter values.

Bayesian uncertainty estimator (2)

Proposed a proof-of-concept noise-robust speaker identification benchmark based on the CHiME domestic noise data.

Results:



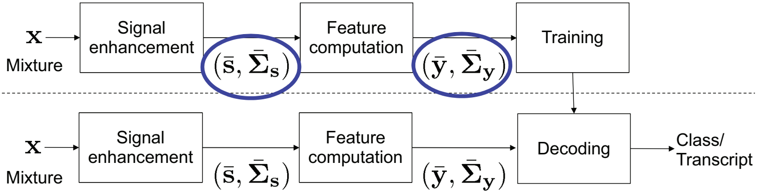
Uncertainty training (1)

Conventional training approaches:

- training on clean data,
- training on noisy data without uncertainty.

Both are biased: the amount of noise is underestimated or overestimated.

Proposed uncertainty training paradigm [postdoc Ozerov]:



Uncertainty training (2)

Derived an EM algorithm that optimizes the uncertainty decoding objective on noisy training data by alternatingly:

- estimating 1st and 2nd order moments of the underlying clean data,
- updating the model parameters given these moments.

% correct on the same robust speaker identification benchmark:

Enhanced signal	Training approach	Decoding approach	Training condition			
			Clean	Matched	Unmatched	Multi
No	Conventional	Conventional	65.17	71.81	69.34	84.09
Yes	Conventional	Conventional	55.22	82.11	80.91	90.12
Yes	Conventional	Uncertainty				
Yes	Uncertainty	Uncertainty				

Uncertainty training (2)

Derived an EM algorithm that optimizes the uncertainty decoding objective on noisy training data by alternatingly:

- estimating 1st and 2nd order moments of the underlying clean data,
- updating the model parameters given these moments.

% correct on the same robust speaker identification benchmark:

Enhanced signal	Training approach	Decoding approach	Training condition: Clean	Matched	Unmatched	Multi
No	Conventional	Conventional	65.17	71.81	69.34	84.09
Yes	Conventional	Conventional	55.22	82.11	80.91	90.12
Yes	Conventional	Uncertainty	**75.51**	78.60	77.58	85.02
Yes	Uncertainty	Uncertainty				

Uncertainty training (2)

Derived an EM algorithm that optimizes the uncertainty decoding objective on noisy training data by alternatingly:

- estimating 1st and 2nd order moments of the underlying clean data,
- updating the model parameters given these moments.

% correct on the same robust speaker identification benchmark:

Enhanced signal	Training approach	Decoding approach	Training condition: Clean	Matched	Unmatched	Multi
No	Conventional	Conventional	65.17	71.81	69.34	84.09
Yes	Conventional	Conventional	55.22	82.11	80.91	90.12
Yes	Conventional	Uncertainty	**75.51**	78.60	77.58	85.02
Yes	Uncertainty	Uncertainty	**75.51**	**82.87**	**81.52**	**91.13**

Best results when using both uncertainty decoding and training. Works even for unmatched training data!

Also applied to singer identification [Lagrange et al., 2012].

Music language modeling: an exploratory study

Language modeling is needed to bridge the semantic gap.

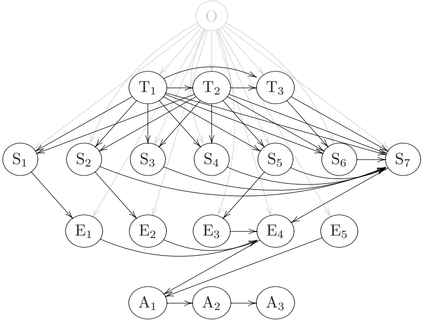
Except a few studies [Raphael, Ryynänen, Mauch...], this issue has been overlooked in music.

Managed Inria EA VERSAMUS project with U. Tokyo.

Roadmap
[Vincent, 2010]

Multiple dependencies
[postdoc Raczyński]

Semiotic structure
[Bimbot, PhD Sargent]



Overall features
O Tags

Temporal features
T_1 Structure
T_2 Meter
T_3 Rhythm

Symbolic features
S_1 Notated tempo
S_2 Notated loudness
S_3 Key/mode
S_4 Harmony
S_5 Instrumentation
S_6 Lyrics
S_7 Quantized notes

Expressive features
E_1 Expressive tempo
E_2 Expressive loudness
E_3 Instrumental timbre
E_4 Expressive notes
E_5 Rendering

Acoustic features
A_1 Tracks
A_2 Mix
A_3 Low-level features

Part 1. Audio source separation

Part 2. Audio content description

Part 3. Research directions

Research directions in source separation

Audio source separation has become a mature topic which is now at the stage of applied research and technology transfer.

Some remaining challenges:

- Benefit from the advantages of both time-domain and Gaussian models
  $\Rightarrow$ unified framework accounting for phase in Gaussian models
- Overcome local optima of the EM algorithms
  $\Rightarrow$ advanced Bayesian inference (structured VB, ensemble models...)
- Address automatic model selection
  $\Rightarrow$ Bayesian model selection
- Deploy real-world applications
  $\Rightarrow$ exploit extra information, e.g., source repetitions [PhD Souviraà].

Research directions in content description

The uncertainty propagation paradigm is still emerging and lies at the frontier of exploratory and applied research.

Some remaining challenges:

- Obtain more accurate and robust uncertainty estimates
  $\Rightarrow$ finer Bayesian approximations (structured VB. . .) [PhD Tran]
- Provide feedback from speech/speaker recognition to source separation
  $\Rightarrow$ constraining spectral envelopes in our flexible spectral model
- Reduce the semantic gap in music processing
  $\Rightarrow$ take the opportunity of the move to PAROLE to exploit and adapt successful approaches in natural language processing [PhD Mesnil].

Conclusion

Mix of short-term and long-term research united by the use and development of a Bayesian modeling and inference framework.

Application focus in PAROLE: speech enhancement and robust speech recognition.

Many more potential applications, including:

- high-fidelity hearing aids and mobile communications,
- voice applications, multimedia document indexing, music search,
- 3D audio rendering, repurposing, interactive applications. . .

Ultimate vision: enhance, understand and interact with complex audio data in a seamless fashion.

Many thanks to. . .

Kamil Adiloğlu	Mathieu Lagrange
Shoko Araki	Stéphanie Lemaile
Roland Badeau	Pierre Leveau
Jon Barker	Jonathan Le Roux
Alexis Benichoux	Dimitris Moreau
Nancy Bertin	Andrew Nesbit
Frédéric Bimbot	Alexey Ozerov
Charles Blandin	Nobutaka Ono
Ngoc Duong	Mark Plumbley
Valentin Emiya	Stanisław Raczyński
Rémi Gribonval	Gabriel Sargent
Nobutaka Ito	Laurent Simon
Maria Jafari	Joachim Thiemann
Matthieu Kowalski	and many others. . .