How to integrate audio source separation and classification?

Emmanuel Vincent

METISS Team
Inria Rennes - Bretagne Atlantique

Includes figures from: R. F. Astudillo, E. Vincent, and L. Deng, “Uncertainty Handling for Environment-Robust Speech Recognition”, Tutorial, Interspeech 2012

1. Introduction
2. The uncertainty handling paradigm
3. Uncertainty estimation
4. Uncertainty propagation
5. Uncertainty decoding and training
6. Conclusion

Audio source separation and classification

Audio signals can be quite complex but source separation has greatly progressed in the last few years!

Music 🔊
↓
Vocals 🔊

Home automation 🔊
↓
Spoken command 🔊

TV series 🔊
↓
Speech 🔊

Audio source separation and classification

Audio signals can be quite complex but source separation has greatly progressed in the last few years!

Music 🔊	Home automation 🔊	TV series 🔊
↓	↓	↓
Vocals 🔊	Spoken command 🔊	Speech 🔊

Can this help audio "classification" tasks such as

- speaker/singer identification,
- speech/lyrics transcription,
- music genre classification, keyword spotting, mutimedia indexing?

Audio source separation: the basics

Audio source separation techniques typically operate in the time-frequency domain, e.g., via the short time Fourier transform (STFT).

They often rely on probabilistic parametric models of the source signals and the mixing process with parameters such as:

- steering/blocking vectors for beamforming, ICA,
- basis spectra and scaling coefficients for NMF, harmonic NMF, and variants thereof,
- exemplar spectra and discrete hidden states for GMM, HMM...

Flexible FASST toolbox integrating several of the above models.
http://bass-db.gforge.inria.fr/fasst/

Audio classification: the basics

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

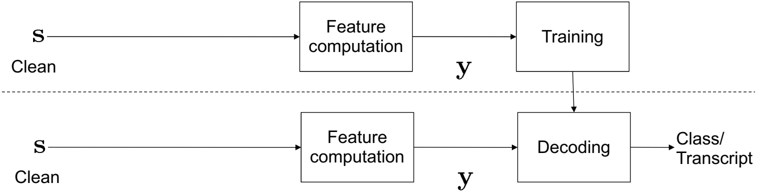
Classification/transcription most often relies on probabilistic acoustic models of the features, e.g., Gaussian mixture models (GMMs).

Audio classification: the basics

Audio content description techniques do not operate in the STFT domain 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.

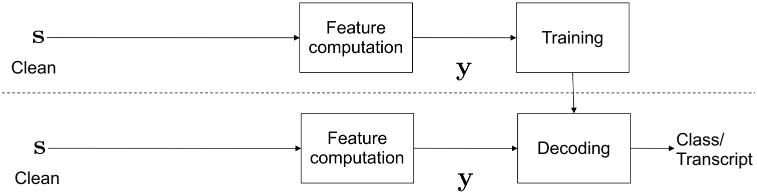


Audio classification: the basics

Audio content description techniques do not operate in the STFT domain 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, doesn't work for noisy data.

How can we reduce or circumvent the mismatch?

Mismatch circumvention techniques

General techniques include:

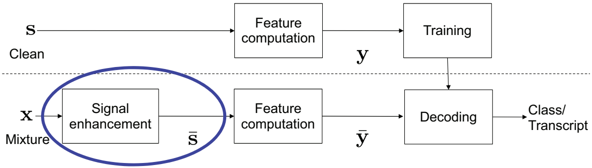
- using different features (modulation features, RNN, etc),
- changing the training objective (discriminative, ARD, etc),
- combinining the outputs of several systems,
- . . .

These non-specific techniques partly circumvent the mismatch between training and test data.

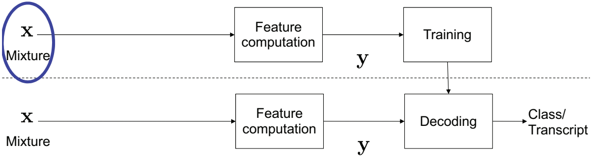
We focus on specific techniques directly aiming to reduce it instead.

Conventional mismatch reduction techniques

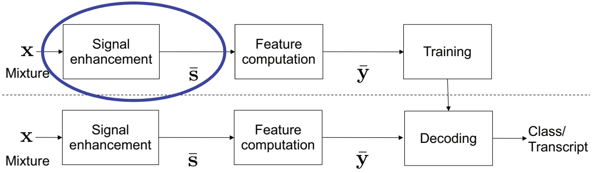
Feature compensation: good separation but often increased mismatch



Training data coverage: better match but huge training set needed



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



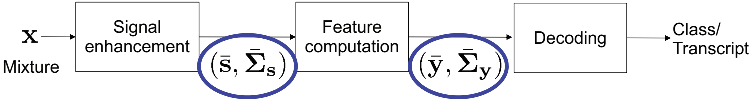
The uncertainty handling paradigm

The uncertainty handling paradigm

Emerging paradigm: estimate and propagate confidence values represented by (approximate) posterior distributions.

Early focus on Boolean uncertainty did not allow complex features.

More recent focus on Gaussian distributions [Deng, Astudillo, Kolossa...].

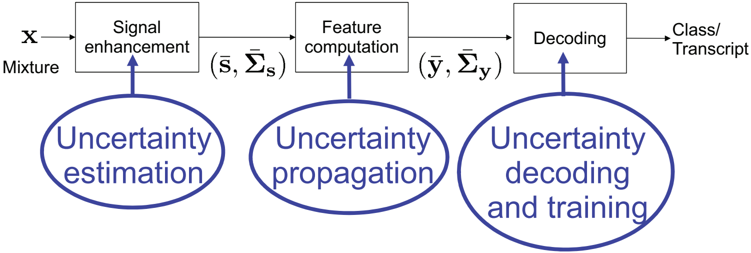

$\mathbf{x}$
Mixture
Signal enhancement
$(\bar{\mathbf{s}}, \bar{\mathbf{\Sigma}}_{\mathbf{s}})$
Feature computation
$(\bar{\mathbf{y}}, \bar{\mathbf{\Sigma}}_{\mathbf{y}})$
Decoding
Class/ Transcript

The uncertainty handling paradigm

Emerging paradigm: estimate and propagate confidence values represented by (approximate) posterior distributions.

Early focus on Boolean uncertainty did not allow complex features.

More recent focus on Gaussian distributions [Deng, Astudillo, Kolossa. . .].


$\mathbf{x}$
Mixture
Signal enhancement
$(\bar{\mathbf{s}}, \bar{\mathbf{\Sigma}}_\mathbf{s})$
Feature computation
$(\bar{\mathbf{y}}, \bar{\mathbf{\Sigma}}_\mathbf{y})$
Decoding
Class/ Transcript
Uncertainty estimation
Uncertainty propagation
Uncertainty decoding and training

Uncertainty estimation

Heuristic uncertainty estimator

First idea: the bigger the change, the bigger the uncertainty [Kolossa].

Variance of $p(\mathbf{s}|\mathbf{x})$ proportional to the squared difference between the noisy signal $\mathbf{x}$ and the separated signal $\bar{\mathbf{s}}$:

$$\bar{\boldsymbol{\Sigma}}_{\mathbf{s},nf} = \mathrm{diag}(\alpha_f |\mathbf{x}_{nf} - \bar{\mathbf{s}}_{nf}|^2).$$

Heuristic solution, not very elegant but somewhat effective.

Wiener uncertainty estimator

Second idea: uncertainty stemming from the Wiener filter [Astudillo].

First estimate the parameters θ of the source models in the maximum likelihood (ML) sense

$$\widehat{\theta} = \arg\max p(\mathbf{x}|\theta).$$

Then, assuming that $\mathbf{x}$ and $\mathbf{s}$ are Gaussian with covariances depending on θ, we have

$$p(\mathbf{s}|\mathbf{x}) \approx p(\mathbf{s}|\mathbf{x}, \widehat{\theta}) = \prod_{nf} \mathcal{N}(\mathbf{s}_{nf}|\bar{\mathbf{s}}_{nf}, \bar{\boldsymbol{\Sigma}}_{\mathbf{s},nf})$$

with

$$\begin{aligned} \bar{\mathbf{s}}_{nf} &= \boldsymbol{\Sigma}_{\mathbf{s},nf} \boldsymbol{\Sigma}_{\mathbf{x},nf}^{-1} \mathbf{x}_{nf} \\ \bar{\boldsymbol{\Sigma}}_{\mathbf{s},nf} &= (\mathbf{I} - \boldsymbol{\Sigma}_{\mathbf{s},nf} \boldsymbol{\Sigma}_{\mathbf{x},nf}^{-1}) \boldsymbol{\Sigma}_{\mathbf{s},nf} \end{aligned}$$

More principled, but does not account for the uncertainty about θ itself!

Bayesian uncertainty estimator (1)

The theoretical Bayesian uncertainty estimator is given by [Adiloğlu]

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

Problem: this integral typically involves thousands of dimensions!

Variational Bayes (VB): approximate the joint posterior $p(\mathbf{s}, \theta|\mathbf{x})$ by the closest distribution $q(\mathbf{s}, \theta)$ for which the integral is tractable.

When q is assumed to factor as $q(\mathbf{s}, \theta) = q(\mathbf{s})q(\theta)$, the posterior over $\mathbf{s}$ is simply obtained as

$$p(\mathbf{s}|\mathbf{x}) \approx q(\mathbf{s}).$$

Bayesian uncertainty estimator (2)

Minimizing the Kullback-Leibler divergence between p and q is equivalent to maximizing the so-called variational free energy

$$\mathscr{L}(q) = \int q(\mathbf{s},\theta) \log \frac{p(\mathbf{x},\mathbf{s},\theta)}{q(\mathbf{s},\theta)} \, d\mathbf{s} \, d\theta$$

This quantity is sometimes not maximizable in closed form: minorization by a parametric bound $f(\mathbf{x},\mathbf{s},\theta,\Omega) \leq p(\mathbf{x},\mathbf{s},\theta)$ may be needed.

Assuming $q(\mathbf{s},\theta) = \prod_{nf} q(\mathbf{s}_{nf}) \prod_i q(\theta_i)$, the solution is iteratively estimated by

1. tightening the bound w.r.t. the variational parameters Ω,
2. $q(\theta_i) \propto \exp[\mathbb{E}_{\mathbf{s},\theta_{i'\neq i}} \log f(\mathbf{x},\mathbf{s},\theta,\Omega)]$
3. $q(\mathbf{s}_{nf}) \propto \exp[\mathbb{E}_{\mathbf{s}_{n'f'\neq nf},\theta} \log f(\mathbf{x},\mathbf{s},\theta,\Omega)]$

Bayesian uncertainty estimator (3)

This results in an expectation-maximization (EM)-like source separation algorithm where posterior distributions over the parameters are updated instead of point estimates as in usual ML-based EM.

Resulting approximating distributions for FASST:

- complex-valued Gaussian for $\mathbf{s}_{nf}$ and for the steering vectors,
- generalized inverse Gaussian for the NMF parameters.

Speaker identification benchmark (1)

CHiME Speech Separation and Recognition Challenge
http://spandh.dcs.shef.ac.uk/chime_challenge/

Data: short spoken commands mixed with genuine noise backgrounds recorded in a family home.

Training: 20 clean utterances from each of 34 speakers
Test: 20 other utterances per speaker, each mixed at 6 different SNRs

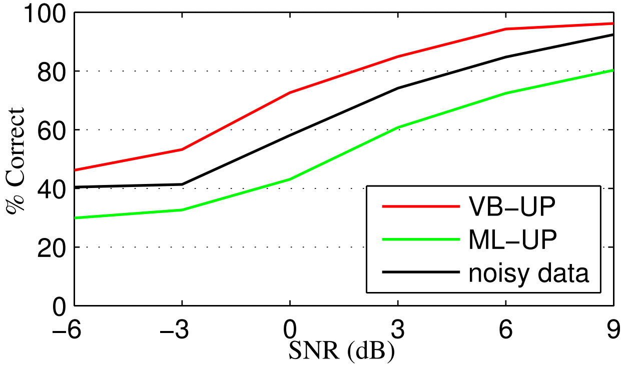
Enhancement: multichannel NMF (ML or VB).

Features: static MFCCs (2 to 20), log-normal uncertainty propagation

Baseline classifier: 32-component GMMs with diagonal covariances, initialized by hierarchical K-means

Speaker identification benchmark (2)

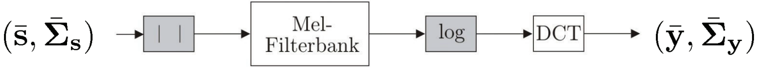
VB performs similarly to ML for the estimation of $\bar{\mathbf{s}}$, but it results in better $\bar{\mathbf{\Sigma}}_\mathbf{s}$ and eventually in improved speaker identification accuracy.



Uncertainty propagation

Mel frequency cepstral coefficients

We take the example of Mel frequency cepstral coefficients (MFCCs):


$(\bar{\mathbf{s}}, \bar{\mathbf{\Sigma}}_{\mathbf{s}})$
$|\ |$
Mel-
Filterbank
log
DCT
$(\bar{\mathbf{y}}, \bar{\mathbf{\Sigma}}_{\mathbf{y}})$

General idea: split the computation into simple steps and propagate the posterior through each step [Gales, Astudillo].

Approximate closed-form equations are preferred to sampling techniques because of their smaller computational cost.

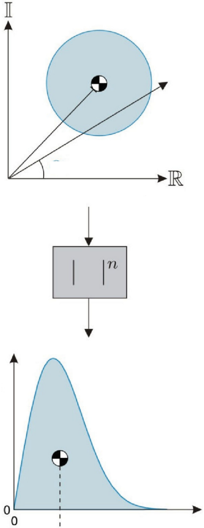
Similar equations can be reused for other features (RASTA-PLP, MLP, chroma, etc).

Propagation through a linear transform

The Mel filterbank and the DCT are linear transforms.

$$\overline{\mathbf{Az}} = \mathbf{A}\bar{\mathbf{z}}$$
$$\bar{\mathbf{\Sigma}}_{\mathbf{Az}} = \mathbf{A}\bar{\mathbf{\Sigma}}_{\mathbf{z}}\mathbf{A}^H$$

Propagation through the magnitude transform



Phase integration results in a Rice distribution whose first and second order moments can be computed in closed form.

$$\overline{|z|} = \bar{\sigma}_z \sqrt{\pi/2}\, L_{1/2}(-|\bar{z}|^2/2\bar{\sigma}_z^2)$$

$$\bar{\sigma}_{|z|}^2 = |\bar{z}|^2 + 2\bar{\sigma}_z^2 - \overline{|z|}^2$$

with $L_{1/2}(z) = e^{z/2}[(1-z)I_0(-z/2) - z\, I_1(-z/2)]$.

Propagation through the log transform

The exponential transform results in a distribution whose first and second order moments can be computed by closed form equations.

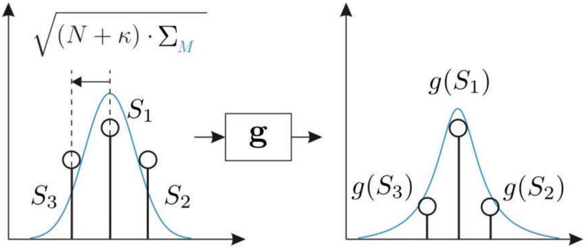
The inversion of these equations yields a good approximation for the logarithmic transform.

$$\overline{\log \mathbf{z}}_i = \log(\bar{\mathbf{z}}_i) - \frac{1}{2}\log\left(\frac{\bar{\boldsymbol{\Sigma}}_{\mathbf{z},ii}}{\bar{\mathbf{z}}_i^2} + 1\right)$$

$$\bar{\boldsymbol{\Sigma}}_{\log \mathbf{z},ij} = \log\left(\frac{\bar{\boldsymbol{\Sigma}}_{\mathbf{z},ij}}{\bar{\mathbf{z}}_i \bar{\mathbf{z}}_j} + 1\right)$$

Propagation through general nonlinear transforms

For general nonlinear transforms, the unscented transform often yields a good approximation at the fraction of the cost of a full sampling scheme.



Vector Taylor series (VTS) has also been proposed.

Uncertainty decoding and training

Gaussian mixture models

We focus for simplicity on Gaussian mixture models (GMMs).

For a given class C, the corresponding GMM is parameterized by

- mean vectors $\boldsymbol{\mu}_i$
- covariance matrices $\boldsymbol{\Sigma}_i$
- weights ω_i.

For clean data, classification is performed in the ML sense: choose the class C that maximizes

$$p(\mathbf{y}|C) = \prod_n \sum_i \omega_i \mathcal{N}(\mathbf{y}_n|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i).$$

The techniques presented here easily extend to hidden Markov models (HMMs) with GMM observation probabilities.

Uncertainty decoding

For noisy data, classification can be performed via the uncertainty decoding (UD) rule [Deng]

$$p(\bar{\mathbf{y}}, \bar{\boldsymbol{\Sigma}}|C) \approx \int p(\mathbf{y}|\bar{\mathbf{y}}, \bar{\boldsymbol{\Sigma}}) p(\mathbf{y}|C)\, d\mathbf{y}$$
$$= \prod_n \sum_i \omega_i \mathcal{N}(\bar{\mathbf{y}}_n|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i + \bar{\boldsymbol{\Sigma}}_n)$$

Model and noise variances add up: this exploits both the model and the uncertainty to (implicitly) predict the missing data.

Modified imputation

An alternative is to replace $\bar{\mathbf{y}}_n$ by its denoised version according to the GMM prior using the Wiener filter:

$$\widehat{\mathbf{y}}_{i,n} = \boldsymbol{\mu}_i + \boldsymbol{\Sigma}_i(\boldsymbol{\Sigma}_i + \bar{\boldsymbol{\Sigma}}_n)^{-1}(\bar{\mathbf{y}}_n - \boldsymbol{\mu}_i)$$

This leads to the heuristic modified imputation (MI) rule [Kolossa]:

$$p(\bar{\mathbf{y}}, \bar{\boldsymbol{\Sigma}}|C) \approx \prod_n \sum_i \omega_i \mathcal{N}(\widehat{\mathbf{y}}_{i,n}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i).$$

Uncertainty training (1)

So far, we have assumed that the GMMs have been trained on clean data but clean data may not be available!

Retraining on noisy data allows to learn the residual distortion that $\bar{\boldsymbol{\Sigma}}_n$ failed to represent because

- the Gaussian parametric model of uncertainty may not fit the actual distribution of uncertainty,
- even when it does, its covariance $\bar{\boldsymbol{\Sigma}}_n$ is never perfectly estimated.

To do so, we maximize the UD criterion over the training data via an EM algorithm considering both the states i_n and the clean data $\mathbf{y}_n$ as hidden data [Ozerov].

Uncertainty training (2)

E-step: estimate clean feature moments by Wiener filtering

$$\begin{aligned}\gamma_{i,n} &\propto \omega_i \mathcal{N}(\bar{\mathbf{y}}_n|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i + \bar{\boldsymbol{\Sigma}}_n),\\ \hat{\mathbf{y}}_{i,n} &= \boldsymbol{\mu}_i + \boldsymbol{\Sigma}_i(\boldsymbol{\Sigma}_i + \bar{\boldsymbol{\Sigma}}_n)^{-1}(\bar{\mathbf{y}}_n - \boldsymbol{\mu}_i),\\ \widehat{\mathbf{R}}_{\mathbf{yy},i,n} &= \hat{\mathbf{y}}_{i,n}\hat{\mathbf{y}}_{i,n}^T + \left(\mathbf{I} - \boldsymbol{\Sigma}_i(\boldsymbol{\Sigma}_i + \bar{\boldsymbol{\Sigma}}_n)^{-1}\right)\boldsymbol{\Sigma}_i.\end{aligned}$$

M-step: update GMM parameters

$$\begin{aligned}\omega_i &= \frac{1}{N}\sum_{n=1}^{N}\gamma_{i,n},\\ \boldsymbol{\mu}_i &= \frac{1}{\sum_{n=1}^{N}\gamma_{i,n}}\sum_{n=1}^{N}\gamma_{i,n}\hat{\mathbf{y}}_{i,n},\\ \boldsymbol{\Sigma}_i &= \mathrm{diag}\left(\frac{1}{\sum_{n=1}^{N}\gamma_{i,n}}\sum_{n=1}^{N}\gamma_{i,n}\widehat{\mathbf{R}}_{\mathbf{yy},i,n} - \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T\right).\end{aligned}$$

Speaker identification benchmark

Same data, source separation algorithm and baseline classifier as above, heuristic STFT uncertainty estimation + log-normal propagation.

Results averaged into 4 training conditions:

- clean,
- matched (same SNR),
- unmatched (different SNR),
- multicondition (all SNRs, hence more noisy data)

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				

Speaker identification benchmark

Same data, source separation algorithm and baseline classifier as above, heuristic STFT uncertainty estimation + log-normal propagation.

Results averaged into 4 training conditions:

- clean,
- matched (same SNR),
- unmatched (different SNR),
- multicondition (all SNRs, hence more noisy data)

Enhanced	Training	Decoding	Training condition			
signal	approach	approach	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				

Speaker identification benchmark

Same data, source separation algorithm and baseline classifier as above, heuristic STFT uncertainty estimation + log-normal propagation.

Results averaged into 4 training conditions:

- clean,
- matched (same SNR),
- unmatched (different SNR),
- multicondition (all SNRs, hence more noisy data)

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**

Singer identification benchmark

Data: 40 songs by 10 singers (5 male and 5 female) from the RWC Popular Music Database, split into 10 s segments

Training/testing: data organized into 4 training/testing folds

Enhancement: melody separation algorithm by Durrieu et al.

Features: static MFCCs (2 to 20), VTS propagation

Baseline classifier: 32-component GMMs with diagonal covariances

Accuracy (%)	per 10 s singing segment					per song (maj. vote)	
	Fold 1	Fold 2	Fold 3	Fold 4	Total	all seg.	sung seg.
no separation	51	53	55	38	49	57	64
wo/ uncertainty	60	63	53	43	55	57	64
w/ uncertainty	**71**	**72**	**84**	**83**	**77**	**85**	**94**

Conclusion

Conclusion

Uncertainty handling is a promising approach for the integration of source separation and classification, which almost reaches the performance achieved on clean data when the uncertainty is known.

Principled uncertainty estimation, propagation and decoding techniques do not always work better than heuristic techniques though. Greater understanding of these heuristics is needed.

Perspectives:

- exploitation the uncertainties about other parameters, e.g., the source spatial location,
- understanding the interplay between source separation methods and uncertainty estimators on the classification performance

Challenges, workshop and resources

2nd CHiME Speech Separation and Recognition Challenge
http://spandh.dcs.shef.ac.uk/chime_challenge/
Deadline: January 15, 2013

4th Signal Separation Evaluation Campaign (SiSEC)
http://sisec.wiki.irisa.fr/
Deadline: Spring 2013, to be announced soon

2nd Int. Workshop on Machine Listening in Multisource Environments
http://spandh.dcs.shef.ac.uk/chime_workshop/
Vancouver, June 1, 2013 (in conjunction with ICASSP)

Other resources
http://lvacentral.inria.fr/
lvalist@googlegroups.com
machinelistening@googlegroups.com

References

K. Adiloğlu and E. Vincent, “A general variational Bayesian framework for robust feature extraction in multisource recordings”, in *Proc. ICASSP*, pp. 273–276, 2012.

R. F. Astudillo and D. Kolossa, “Uncertainty propagation,” in *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*, D. Kolossa and R. Haeb-Umbach, Eds., pp. 35–64. Springer, 2011.

L. Deng, J. Droppo, and A. Acero, “Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion,” *IEEE Trans. on Speech and Audio Processing*, 13(3):412–421, 2005.

M. J. F. Gales, *Model-based techniques for noise robust speech recognition*, PhD thesis, University of Cambridge, 1995.

D. Kolossa, A. Klimas, and R. Orglmeister, “Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques,” in *Proc. WASPAA*, pp. 82–85, 005.

M. Lagrange, A. Ozerov, and E. Vincent, “Robust singer identification in polyphonic music using melody enhancement and uncertainty-based learning”, in *Proc. 13th Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2012.

A. Ozerov, M. Lagrange, and E. Vincent, “Uncertainty-based learning of acoustic models from noisy data”, *Computer Speech and Language*, to appear.