VARIATIONAL EM FOR CLUSTERING INTERAURAL PHASE CUES IN MESSL FOR BLIND SOURCE SEPARATION OF SPEECH

Zeinab Zohny, Syed Mohsen Naqvi, Jonathon A. Chambers

Advanced Signal Processing Group, School of Electronic, Electrical, and Systems Engineering
Loughborough University, Leicestershire, UK
{z.zohny, s.m.r.naqvi, j.a.chambers}@lboro.ac.uk

ABSTRACT

The model-based expectation maximization source separation and localization (MESSL) technique is a probabilistic time-frequency masking algorithm that achieves underdetermined blind source separation of speech sources. Using only two-channel recordings, MESSL clusters spectrogram points based on their interaural spatial cues. Gaussian mixture models (GMMs) are assumed for the interaural cues and their corresponding parameters are determined by maximum likelihood estimation (MLE) via the expectation maximization (EM) framework. However, the presence of singularities and over-fitting are major drawbacks of MLE. In this paper, we investigate variational Bayesian (VB) inference for clustering spectrogram points based particularly on their interaural phase difference (IPD) cues. Variational inference overcomes the difficulties associated with the likelihood optimization and improves the separation especially when the sources are in close proximity. Simulation studies based on speech mixtures formed from the TIMIT database confirm the advantage of the proposed approach in terms of signal to distortion ratio (SDR).

Index Terms— Blind source separation, time-frequency masking, Gaussian mixture models, expectation-maximization, variational Bayesian inference;

1. INTRODUCTION

Amazingly, humans manage to selectively recognize the utterance of one speaker usually in the presence of other interfering speakers, background noise and music. The cocktail party problem (CPP) proposed by Colin Cherry [1] refers to this psychoacoustic phenomenon. For the last decades, numerous efforts [2] were dedicated to unveil the mystery of the human auditory perception capability especially with the growing number of applications requiring speech-based human machine interfaces (HMIs). Many models in the field of blind source separation (BSS) have emerged to tackle the cocktail party problem such as independent component analysis (ICA) [3] and independent vector analysis (IVA) [4]. These approaches exploit the statistical independence of the speech sources to achieve frequency domain convolutive blind source separation (FDCBSS) basically using linear transformations. Linear filtering commonly involving mixing matrix pseudo-inversion works well only when the number of sensors (microphones) is greater or equal to the number of sources [5]. However, when the number of sensors is less than the number of sources (underdetermined or overcomplete case), direct estimation of sources becomes more appropriate and is usually achieved by non-linear techniques such as time-frequency (T-F) masking [6] or the line orientation separation technique (LOST) [7].

Originated in the field of computational auditory scene analysis (CASA) [8], T-F masking approaches exploit the sparseness that acoustic signals exhibit in the time-frequency representation. This property referred to as W-disjoint orthogonality [9] assumes that most of the energy at each T-F or spectrogram point belongs to a single source. MESSL described in [6] is a probabilistic T-F masking based technique that separates multiple sound sources from only two channel recordings in the presence of reverberation and noise. It combines the interaural spatial cues that humans use for localizing with the missing data approach [10] for speech recognition. The interaural cues of each speech source at each T-F point are independently modelled using Gaussian mixture models (GMMs). The parameters of the models and the regions that best fit each model are evaluated using the expectation maximization (EM) algorithm and as a by-product MESSL generates soft probabilistic spectrogram masks for separating individual speech sources. In MESSL, localization is used to initialize the algorithm and the separation performance is mostly dependent on the modelling of the interaural cues and the clustering framework.

In [11], [12] we exploited an alternative modelling of the interaural cues based on the Student's t-distribution and the GMMs were replaced by Student's t-distribution mixture models (SMMs). Due to its heavy tail behaviour, the Student's t-distribution is known to be less sensitive to outlier values and experimental results have confirmed a significant improvement in the average separation performance through the resulting non-Gaussian based robust clustering. In this

paper, we propose the use of VB inference as an alternative to MLE for the same GMMs employed in MESSL. The paper is organized as follows: In Section 2, the EM framework employed in MESSL for clustering IPD GMMs is introduced and the limitations of the likelihood optimization are explained. In Section 3, the variational inference is thoroughly described. Experimental results are shown in Section 4 and finally the relation to prior work is further discussed in Section 5.

2. IPD GAUSSIAN MIXTURE MODELS

MESSL attempts to mimic the sound separation abilities of the human auditory system and hence uses similar cues for localizing sound sources such as the interaural time, phase and level differences. Following [9], we assume that $L(\omega,t)$ and $R(\omega,t)$ are the spectrograms of the mixture signals arriving at two spatially distinct microphones and the interaural spectrogram can be expressed as

$$\frac{L(\omega,t)}{R(\omega,t)} = 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)} \tag{1}$$

where $\phi(\omega,t)$ and $\alpha(\omega,t)$ denote the interaural phase difference (IPD) and the interaural level difference (ILD) measured in dB, respectively. ILD results from the shadowing of the far ear by the head for sounds typically above 3-4 kHz. On the other hand, IPD conveys information about the azimuthal location of a sound source. In order to avoid ambiguities and phase circularity [9], the phase residual $\hat{\phi}(\omega,t;\tau)$ expressed as

$$\hat{\phi}(\omega,t;\tau) = arg\Big(e^{j\phi(\omega,t)} e^{-j\omega\tau(\omega)}\Big) \tag{2}$$

is used instead of $\phi(\omega,t)$. It can be modelled approximately by a Gaussian distribution $\mathcal{N}(\hat{\phi}(\omega,t;\tau)|\xi(\omega),\sigma^2(\omega))$ with mean $\xi(\omega)$ and variance $\sigma^2(\omega)$ [9]. Based on the W-disjoint orthogonality [9], each spectrogram point belongs only to a source i and delay $\tau(w)$. The delay is expressed as

$$\tau(\omega) = \tau + \omega^{-1}\xi(\omega) \tag{3}$$

where τ is a discrete random variable used for localization while the parameter $\xi(\omega)$ is varying randomly with frequency in the interval $(-\pi,\pi)$. Although the number of sources is assumed known, the source i dominating each spectrogram point, as well as the delay τ, are latent variables. Both hidden variables can however be combined into one latent variable $z_{i\tau}(w,t)$. This parameter is equal to one with a corresponding probability $\psi_{i\tau}$, if the spectrogram point belongs to source i and delay τ and zero otherwise. In other words, $z_{i\tau}(\omega,t) \in \{0,1\}$ and $\sum_{i,\tau} z_{i\tau}(\omega,t) = 1$. Let $\Theta \equiv \{\xi_{i\tau}(\omega), \sigma^2_{i\tau}(\omega), \psi_{i\tau}\}$ denote the set of the parameters of the models. The likelihood for a given observation can be expressed as

$$\mathcal{L}(\Theta) = \sum_{\omega,t} \log \sum_{i,\tau} [\mathcal{N}(\hat{\phi}(\omega,t;\tau)|\xi_{i\tau}(\omega), \sigma^2_{i\tau}(\omega).\psi_{i\tau}] \tag{4}$$

The above equation represents the log-likelihood of the GMMs with one Gaussian per (i,τ) combination and $\psi_{i\tau}$ as the mixing weights. The parameters of each model can be determined using the iterative EM algorithm involving two steps. In the E step, the expectations of the latent variable $z_{i\tau}(\omega,t)$ denoted by $r_{i\tau}(\omega,t)$ are computed given the current observations and the parameters estimates. These values are then used in the M step to maximize the log-likelihood and re-estimate Θ. In addition to the model parameters, MESSL generates probabilistic masks for each of the sources expressed as

$$M_i(\omega,t) \equiv \sum_{\tau} r_{i\tau}(\omega,t) \tag{5}$$

The major problem with the EM algorithm is the potential unbounded property of the likelihood [19]. In other words, if one component has its mean exactly equal to one of the data points, its contribution to the likehood function can be written as

$$\mathcal{N}(\hat{\phi}(\omega,t)|\xi_{i\tau}(\omega), \sigma^2_{i\tau}(\omega)) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma_{i\tau}(\omega)} \tag{6}$$

as $\sigma_{i\tau}(\omega)$ tends to 0, the likelihood function tends to infinity. These singularities will always occur whenever one of the Gaussian components collapses onto a data point. Detection of such singularities and avoiding them is crucial when adopting MLE [13]. This difficulty does not occur if a VB approach is employed.

3. VARIATIONAL INFERENCE FOR GAUSSIAN MIXTURE MODELS

In contrast to MLE, in a VB approach, the parameters are also treated as random variables and prior distributions are imposed on these parameters. For each observation $\hat{\phi}(\omega,t;\tau)$, there is corresponding binary vector $\mathbf{z}(\omega,t)$ comprising the elements $z_{i\tau}(\omega,t)$, and the prior distribution of $\mathbf{Z}$ given the mixing weights can be written in the form

$$p(\mathbf{Z}|\psi) = \prod_{t} \prod_{i,\tau} \psi_{i\tau}(\omega)^{z_{i\tau}(\omega,t)} \tag{7}$$

where $\mathbf{Z} = \{\mathbf{z}(\omega,t)\}$ and $\boldsymbol{\psi} = \{\psi_{i\tau}(\omega)\}$. The number of the latent variables $z(\omega,t)$ increases with the size of the data set. However, the size of the parameters is fixed independent of the data size. For analytical simplicity, conjugate prior distributions are considered for modelling the parameters [13]. Hence, at each frequency ψ can be modelled by the Dirichlet density

$$p(\boldsymbol{\psi}) = Dir(\boldsymbol{\psi}|\alpha_0) = C(\alpha_0) \prod_{i\tau} \psi_{i\tau}^{\alpha_0 - 1} \tag{8}$$

where α_0 is the distribution parameter assumed to be the same for all components and $C(\alpha_0)$ is the normalization constant. At each frequency, the mean and the precision parameters are

modelled by a Gaussian-Wishart prior (namely the conjugate of a Gaussian distribution) given by

$$p(\boldsymbol{\xi}, \boldsymbol{\lambda}) = p(\boldsymbol{\xi}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}) \tag{9}$$

$$= \prod_{i\tau} \mathcal{N}(\xi_{i\tau}|m_0, (\beta_0\lambda_{i\tau})^{-1}\mathcal{W}(\lambda_{i\tau}|w_0, \nu_0) \tag{10}$$

where $\lambda_{i\tau}$ denotes the inverse of the variance, $\boldsymbol{\xi} = \{\xi_{i\tau}\}$, $\boldsymbol{\lambda} = \{\lambda_{i\tau}\}$ and m_0, β_0, w_0, ν_0 are the Gaussian-Wishart distribution parameters [13]. m_0 is chosen to be equal to the mean of the data [14] and hence is frequency dependent whereas β_0, w_0 and ν_0 are frequency independent and fixed a priori. All the Gaussian-Wishart hyperparameters are assumed equal for all components. Our goal is the estimation of the posterior distributions of all the hidden variables given the data set. This would be approximated by a distribution $q^*(\mathbf{Z}, \boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{\lambda})$ minimizing the Kullback-Leibler divergence functional [13] and satisfying the only assumption of the variational inference, namely

$$q^*(\mathbf{Z}, \boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{\lambda}) = q^*(\mathbf{Z})q^*(\boldsymbol{\psi})q^*(\boldsymbol{\xi}, \boldsymbol{\lambda}) \tag{11}$$

The optimal distributions $q^*(\mathbf{Z})$, $q^*(\boldsymbol{\psi})$ and $q^*(\boldsymbol{\xi}, \boldsymbol{\lambda})$ have the same functional form as their priors [13]. Similarly to the EM algorithm, these variational posterior distributions are obtained in two steps. In the E-step, the current distributions are used to evaluate $E[z_{i\tau}(\omega, t)]$ followed by the M-step in which the parameters of the distributions are recomputed given the expected value of $z_{i\tau}(\omega, t)$. Within the E-step, the expected value $r_{i\tau}(\omega, t)$ is computed as follows

$$r_{i\tau}(\omega, t) = \frac{\rho_{i\tau}(\omega, t)}{\sum_{i\tau} \rho_{i\tau}(\omega, t)} \tag{12}$$

where

$$\ln \rho_{i\tau}(\omega, t) = E[\ln \psi_{i\tau}(\omega)] + \frac{1}{2}E[\ln \lambda_{i\tau}(\omega)] - \frac{1}{2}\ln(2\pi) - \frac{1}{2}E_{\xi_{i\tau},\lambda_{i\tau}}[\big(\hat{\phi}(\omega, t; \tau) - \xi_{i\tau}(\omega)\big)^2] \tag{13}$$

The following three statistics related to $r_{i\tau}(\omega, t)$ are defined as

$$N_{i\tau}(\omega) = \sum_t r_{i\tau}(\omega, t) \tag{14}$$

$$\bar{\phi}_{i\tau}(\omega) = \frac{1}{N_{i\tau}(\omega)} \sum_t \hat{\phi}(\omega, t; \tau) r_{i\tau}(\omega, t) \tag{15}$$

$$S_{i\tau(\omega)} = \frac{1}{N_{i\tau}(\omega)} \sum_t \big((\hat{\phi}(\omega, t; \tau) - \bar{\phi}_{i\tau}(\omega))^2 r_{i\tau}(\omega, t)\big) \tag{16}$$

and are used in the evaluation of the parameters of the variational posterior distributions in the M step as follows,

$$\alpha_{i\tau}(\omega) = \alpha_0 + N_{i\tau}(\omega) \tag{17}$$

$$\beta_{i\tau}(\omega) = \beta_0 + N_{i\tau}(\omega) \tag{18}$$

$$m_{i\tau}(\omega) = \frac{1}{\beta_{i\tau}(\omega)}(\beta_0 m_0(w) + N_k\bar{\phi}_{i\tau}(\omega)) \tag{19}$$

$$w_{i\tau}(\omega)^{-1} = w_0^{-1} + N_{i\tau}(\omega)S_{i\tau}(\omega) + \frac{\beta_0 N_{i\tau}(\omega)}{\beta_0 + N_{i\tau}(\omega)}(\bar{\phi}_{i\tau}(\omega) - m_0)^2 \tag{20}$$

$$\nu_{i\tau}(\omega) = \nu_0 + N_{i\tau}(\omega) \tag{21}$$

where $\alpha_{i\tau}(\omega)$ is the parameter of the updated Dirichlet distibution $q^*(\psi)$ and $\beta_{i\tau}(\omega)$, $m_{i\tau}(\omega)$, $w_{i\tau}(\omega)$ and $\nu_{i\tau}(\omega)$ define the parameters of the updated Gaussian-Wishart distribution.

These parameters are then used to compute the set of expectations $E[\ln \psi_{i\tau}(\omega)]$, $E[\ln \lambda_{i\tau}(\omega)]$ and $E_{\xi_{i\tau},\lambda_{i\tau}}[(\hat{\phi}(\omega, t; \tau) - \xi_{i\tau}(\omega))^2]$ required for estimating $r_{i\tau}(\omega, t)$

$$E_{\xi_{i\tau},\lambda_{i\tau}}[\big(\hat{\phi}(\omega, t; \tau) - \xi_{i\tau}(\omega)\big)^2] = 1/\beta_{i\tau}(\omega) + \nu_{i\tau}(\omega)\Big(\hat{\phi}(\omega, t; \tau) - m_{i\tau}(\omega)\Big)^2 w_{i\tau}(\omega) \tag{22}$$

$$E[\ln \lambda_{i\tau}(\omega)] = \psi\big(\frac{\nu_{i\tau}(\omega)}{2}\big) + \ln 2 + \ln w_{i\tau}(\omega) \tag{23}$$

$$E[\ln \psi_{i\tau}(\omega)] = \psi(\alpha_{i\tau}(\omega)) - \psi\Big(\sum_{i\tau} \alpha_{i\tau}(\omega)\Big) \tag{24}$$

where $\psi(.)$ is the digamma function [13]. After convergence, the values $r_{i\tau}(w, t)$ are used to compute the masks as indicated in equation (5). Initialization of $r_{i\tau}(w, t)$ follows MESSL [9]. Estimates of τ for each source are determined using the Phase Transform (PHAT) [15]. $\psi_{i\tau}$ is then assumed initially to have a Gaussian distribution with its mean located at each cross correlation maximum and a standard deviation of one sample. The first E-step is calculated assuming zero means and unit variances followed by the M-step, these two steps are repeated until convergence.

4. EXPERIMENTAL RESULTS

In these simulations, we randomly chose different speech signals from the whole TIMIT database [16]. Each signal is 2.5 s long. These signals were normalized and convolved with real binaural impulse responses recorded at a reverberation time RT60$\approx$ 565 ms [17]. The sampling frequency was 8kHz. The target was always directed in the front of the microphones and since we are interested in the case where sources are in close proximity, three different azimuthal positions for the interferer were tested [15°, 30°, 45°], in the case of two speakers. In the three-speaker case, the second interferer is located symmetrically with the same azimuth. All speech sources are located at a distance of 1 m from the center of the microphones. The separation performance was evaluated objectively by the signal-to-distortion ratio (SDR) [18]. Let Θ_Ω denote the complexity in which IPD parameters vary with the frequency. In

MESSL, this complexity results in a better separation than the frequency independent version but requires a bootstrapping approach to avoid local maxima [9]. Our approach also assumes frequency dependent parameters with less complexity as no bootstrapping is required [19]. The set of hyperparameters can be fixed a priori or can be inferred from the data. In our experiments, β_0 and $m_0(\omega)$ were set following [14], where $\beta_0 = 0.01$, $m_0(\omega)$ is equal to the mean of the data at each frequency and ν_0 was chosen empirically equal to 20 as smaller values resulted in slower convergence. The Dirichlet distribution hyperparameter α_0 plays an important role in variational clustering as it can be seen as the effective prior number of observations associated with each component [13]. Solutions obtained for $\alpha_0 < 1$ correspond to the case where more mixing coefficients are equal to zero which better describes our problem. Individual SDRs obtained for five mixtures using our proposed approach with different values of α_0 are shown in Table 1 and Table 2. Poor choice of prior distribution might affect the effectiveness of the VB approach as indicated in Table 2, where $\alpha_0 = 10$ and the average SDRs have been reduced by 0.9 dB, 1.1 dB and 1.2 dB for the three azimuthal separation angles respectively. We randomly formed

Table 1: SDR (dB) proposed approach Θ_Ω, $\alpha_0 = 0.1$

Azimuth angles	15°	30°	45°
mix1	2.53	3.49	2.79
mix2	3.57	3.1	5.1
mix3	3.55	3.99	4.95
mix4	3.16	3.08	3.3
mix5	2.98	2.17	3.35
Average	3.16	3.16	3.9

Table 2: SDR (dB) proposed approach Θ_Ω, $\alpha_0 = 10$

Azimuth angles	15°	30°	45°
mix1	1.43	2.62	1.5
mix2	2.62	2.37	4.29
mix3	2.9	2.99	4.04
mix4	1.97	1.09	1.54
mix5	2.31	1.01	2.05
Average	2.24	2.01	2.68

10 different mixtures in total from the TIMIT database and the average SDR results comparing our approach ($\alpha_0 = 0.1$) with two versions of MESSL are shown in Table 3 and Table 4 for two and three speakers, respectively. It can be seen that adding ILD cues for small separation angles does not improve the separation which is expected since both spatial cues get more similar as the sources move closer [9]. On the other hand, exploiting VB clustering framework improves the estimation of the parameters of IPD cues for sources in close proximity, resulting in more accurate masks and a better separation. The average SDR improvement of the proposed approach decreases with the azimuthal separation. For the two-speaker case, the average SDR improvements obtained using the variational approach are 1.2 dB, 0.8 dB and 0.5 dB compared to the first version of MESSL. Whereas, compared to the second version MESSL IPD-ILD, the average SDR improvements obtained are 1.7 dB, 1.2 dB and 0.7 dB for the three azimuthal angles respectively. In Table 4, for the case of three speakers these improvements increased to 1.5 dB, 1.1 dB and 0.9 dB compared to the first version and 1.9 dB, 1.2 dB and 0.8 dB compared to the second version.

Table 3: Separation performance comparison in terms of average SDR (dB) for the two-speaker case

Azimuth angles	15°	30°	45°
MESSL IPD	2.38	2.62	3.67
MESSL IPD-ILD	1.92	2.22	3.47
Variational IPD	3.61	3.42	4.13

Table 4: Separation performance comparison in terms of average SDR (dB) for the three-speaker case

Azimuth angles	15°	30°	45°
MESSL IPD	-0.67	0.09	2.11
MESSL IPD-ILD	-1.15	-0.02	2.22
Variational IPD	0.8	1.22	2.97

5. RELATION TO PRIOR WORK

The state-of-the-art model-based expectation maximization source separation and localization (MESSL) algorithm separates successfully multiple sound sources from only two-channel reverberant mixtures. In this paper, we improved the MESSL clustering framework by exploiting variational Bayesian inference as an alternative to the likelihood maximization approach. The proposed framework is used for clustering spectrogram points based on their interaural phase difference (IPD) cues. This elegant approach overcomes the drawbacks of the popular EM for GMMs as it avoids overfitting and the presence of singularities associated with the likelihood optimization without requiring additional extensive computations. More importantly, with proper initialization and careful choice of hyperparameters values, experimental results confirmed an improvement of the separation performance particularly for small separation angles. Future work will consider integrating the robust clustering resulting from the non-Gaussian modelling within the variational Bayesian framework to cluster the spectrogram points based on both interaural phase and level difference cues. The variational approach might also be used to determine the number of active speech sources.

6. REFERENCES

[1] C. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears," *The Journal of The Acoustical Society of America*, vol. 25, no.5, pp.975-979, 1953.

[2] S. Haykin and Z. Chen, "The Cocktail Party Problem," *Neural Computation*, 17, pp.1875-1902, 2005

[3] A. Hyvarinen and E. Oja, "Independent Component Analysis: Algorithms and Applications," *Neural Networks*, Volume 13, 4-5, pp. 411-430, 2000.

[4] T. Kim, I. Lee, S. Lee and T. Lee, "Independent Vector Analysis: Definition and Algorithms", *Signals, Systems and Computers*, pp.1393-1396, 2006.

[5] P.D. O'Grady, B.A Pearlmutter and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," *International Journal of Imaging Sys. and Tech.*, vol 15, pp. 18-33, 2005.

[6] M. I. Mandel, *Binaural Model-Based Source Separation and Localization*, PhD thesis, Columbia University, 2010.

[7] P.D. O'Grady and B.A Pearlmutter, "Soft-LOST: EM on a mixture of Oriented Lines," in *Proc. ICA (LNCS 3195)*, pp. 430-436, 2004.

[8] D. Wang, "Time-frequency Masking for Speech Separation and its potential for Hearing Aid Design," *Trends in Amplification*, vol. 12, no.4, pp. 332-351, 2008.

[9] M. I. Mandel, R. J. Weiss and D. P. W. Ellis, "Model-Based Expectation-Maximization Source Separation and Localization," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no.2, pp. 382-394, 2010.

[10] S. Harding, J. Barker and G. J. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no.1, pp. 58-67, 2006.

[11] Z. Zohny, S. M. Naqvi and J. A. Chambers, "Enhancing the MESSL algorithm with robust clustering based on the Student's t-distribution," *IEEE Electron. Lett.*, vol. 50, issue 7, 2014.

[12] Z. Zohny and J. A. Chambers, "Modelling interaural level and phase cues with Student's t-distribution for robust clustering in MESSL," in *Proc. Dig. Sig. Process.*, pp. 59-62, 2014.

[13] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[14] C. Fraley and A. E. Raftery "Bayesian regularization for normal mixture estimation and model-based clustering," *Journal of Classification*, vol. 24, pp. 151-181, 2007.

[15] P. Aarabi, "Self-Localizing Dynamic Microphone Arrays," *IEEE Trans. Syst., Man, Cybern. Part C*, vol. 32, no. 4, pp. 474-484, 2002.

[16] J. S. Garofolo et al., 'TIMIT Acoustic-Phonetic Continuous Speech Corpus', *Linguistic Data Consortium*, 1993.

[17] B. S. Cunningham, N.Kopco and T.Martin,'Localizing nearby sound sources in a classroom: Binaural room impulse responses', *J. Acoust. Soc. Amer.* pp. 3100-3115, 2005.

[18] E. Vincent, C. Fevotte and R. Gribonval, "Performance measurement in Blind Audio Source Separation," *IEEE Trans. Speech and Audio Processing*, vol. 14, No. 4, pp. 1462-1469, 2006.

[19] C. Archambeau and M. Verleysen, "Robust Bayesian clustering," *Neural Networks*, pp. 127-138, 2007.