4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), April 2003, Nara, Japan

WIENER BASED SOURCE SEPARATION WITH HMM/GMM USING A SINGLE SENSOR

Laurent BENAROYA, Frédéric BIMBOT

IRISA (CNRS & INRIA), METISS, Campus de Beaulieu
35042 Rennes Cedex, France

ABSTRACT

We propose a new method to perform the separation of two audio sources from a single sensor. This method generalizes the Wiener filtering with Gaussian Mixture distributions and with Hidden Markov Models. The method involves a training phase of the models parameters, which is done with the classical EM algorithm. We derive a new algorithm for the re-estimation of the sources with these mixture models, during the separation phase. The general approach is evaluated on the separation of real audio data and compared to classical Wiener filtering.

1. INTRODUCTION

We use in this paper the Wiener filter theory for source separation using one single sensor.
We observe a mixture $x(t) = s_1(t) + s_2(t)$ which is the superimposition of two sources and our aim is to estimate the sources. In the Bayesian context, it is assumed in standard Wiener filtering that the prior densities of the sources are gaussian (centered) densities. Some attempts have been made, in the denoising context, to use other prior densities for the original signal model : generalized gaussian densities [1] or gaussian mixture priors [2] and gaussian scale mixture models [3], [4].

We study in this paper gaussian mixture priors (GMM) as well as Hidden Markov Models with gaussian conditional densities, in the context of source separation. This involves two non gaussian models, one for each source, the parameters of which are estimated in a training phase. We derive a source estimation algorithm, which takes the form of an adaptive weighted Wiener filtering, in the separation phase.

Tgis work is on a similar line as [5] but it generalises the approach to GMM models and addresses the case of smooth adaptive Wiener filtering

As we are interested in this paper in audio sources, or more generally, locally stationary sources, we will not study the signals directly, but rather their Short Term Fourier Transform (STFT), denoted by $\mathcal{S}$. The mixing equation becomes $\mathcal{S}x(t,f) = \mathcal{S}s_1(t,f) + \mathcal{S}s_2(t,f)$, in the time-frequency domain, as the STFT is linear.

If we assume that both sources have gaussian centered priors, with respective diagonal covariance matrices $\Sigma_1 = \text{diag}(\sigma_1^2(f))$ and $\Sigma_2 = \text{diag}(\sigma_2^2(f))$, then the optimal Bayesian estimators for both sources is the Wiener filtering [6] :

$$\widehat{\mathcal{S}s}_1(t,f) = \frac{\sigma_1^2(f)}{\sigma_1^2(f)+\sigma_2^2(f)}\mathcal{S}x(t,f)$$
$$\widehat{\mathcal{S}s}_2(t,f) = \frac{\sigma_2^2(f)}{\sigma_1^2(f)+\sigma_2^2(f)}\mathcal{S}x(t,f)$$

In this paper, we study the case of mixture of gaussian priors, for the STFT of both sources

$$p_i(\mathcal{S}s_i(t,f)) = \sum_{k_i=1}^{Q_i} \omega_{k_i} p_G(\mathcal{S}s_i(t,f); \{\sigma_{k_i}^2(f)\})$$

Where Q_i is number of components in the model of the source s_i and $p_G(\mathcal{S}s(t,f), \{\sigma^2(f)\})$ is the centered gaussian density with covariance matrix $\text{diag}(\sigma^2(f))$ and $\sigma_{k_i}^2(f)$ is the variance of the gaussian density corresponding to the index k_i of the source s_i ($i = 1, 2$), and to the frequency component f.

$$p_G(\mathcal{S}s_i(t,f), \{\sigma_{k_i}^2(f)\}) = \frac{1}{(2\pi)^{d/2}\prod_f \sigma_{k_i}(f)} \times \exp\left[-0.5\sum_f \frac{|\mathcal{S}s_i(t,f)|^2}{\sigma_{k_i}^2(f)}\right]$$

The diagonal covariance matrices $\text{diag}(\sigma_{k_i}^2(f))$ can be interpreted as power spectral densities (PSD) which correspond to spectral shapes. These spectral shapes correspond to the structure of audio signals which contain various types of timbers and pitches. Our motivation for the use of Gaussian Mixture Models (GMM) is to take into account the diverse structure of sounds through multiple PSD (covariance matrices). Thus we may model non stationary, still locally stationary, signals contrary to classical Wiener filters.

In fact, this prior structure is quite general and leads to an adaptive Wiener filtering scheme.

It is assumed here that both Gaussian Mixture Models (GMM) parameters are estimated in a first phase, in which

training samples of the sources are provided in order to estimate the covariance matrices and the prior weights.

In section 2, we propose an algorithm for the estimation of the sources, in the separation phase, in the case of gaussian mixture priors for the sources (GMM).
In section 3, we discuss the training phase, which relies on the Expectation-Maximization (EM) algorithm.
In section 4, we generalize our prior models to Hidden Markov Models (HMM) with gaussian conditional densities.
In section 5, we study both GMM/HMM models on a mixture of real audio signals.

2. SOURCE ESTIMATION

The aim here is to derive an estimation algorithm of the sources, given the mixture, in the GMM setting. We will see that this estimation can be expressed as weighted Wiener filters, where the weights are adaptive. We use in this paper a classical incomplete data formalism for the Gaussian Mixture Models (GMM).

If $X(t,f)$ is the observed signal (or $\mathcal{S}x(t,f)$), for a given t, we suppose that the complete data is $Z = \{X(t,f), q(t)\}$ where $q(t)$ is the index of the active component in the GMM at time index t, that is the index of the gaussian density from which the data $X(t,f)$ was generated.

If we know, in the additive mixture setting $\mathcal{S}x(t,f) = \mathcal{S}s_1(t,f) + \mathcal{S}s_2(t,f)$, for a given time t, the active component indexes k_1 and k_2 for both sources s_1 and s_2, then the conditional posterior mean estimator coincides with the Wiener estimator :

$$E[\mathcal{S}s_1(t,f)|(k_1,k_2)] = \frac{\sigma^2_{k_1}(f)}{\sigma^2_{k_1}(f)+\sigma^2_{k_2}(f)}\mathcal{S}x(t,f)$$
$$E[\mathcal{S}s_2(t,f)|(k_1,k_2)] = \frac{\sigma^2_{k_2}(f)}{\sigma^2_{k_1}(f)+\sigma^2_{k_2}(f)}\mathcal{S}x(t,f)$$

We now introduce the posterior probability of the components k_1, k_2 of both models of the sources $\mathcal{S}s_1(t,f)$ and $\mathcal{S}s_2(t,f)$ at time t, denoted by $\gamma_{k_1,k_2}(t)$. Thus $\gamma_{k_1,k_2}(t)$ is the probability of the hidden (active) components $q_1(t)$ and $q_2(t)$, at time t, are equal to k_1, k_2 conditionally to the complete observed sequence $\{\mathcal{S}x(t_1,f),\ldots,\mathcal{S}x(t_N,f)\}$:

$$\gamma_{k_1,k_2}(t) = p_t(q_1 = k_1, q_2 = k_2 | \mathcal{S}x(t_1,f),\ldots,\mathcal{S}x(t_N,f))$$

We estimate the sources with the different Wiener filters for each couple of components (k_1,k_2), weigthed with the posterior probability $\gamma_{k_1,k_2}(t)$

$$\widehat{\mathcal{S}s}_1(t,f) = \left[\sum_{k_1,k_2}\gamma_{k_1,k_2}(t)\frac{\sigma^2_{k_1}(f)}{\sigma^2_{k_1}(f)+\sigma^2_{k_2}(f)}\right]\mathcal{S}x(t,f)$$
$$\widehat{\mathcal{S}s}_2(t,f) = \left[\sum_{k_1,k_2}\gamma_{k_1,k_2}(t)\frac{\sigma^2_{k_2}(f)}{\sigma^2_{k_1}(f)+\sigma^2_{k_2}(f)}\right]\mathcal{S}x(t,f)$$

This is a time varying Wiener-type filter. It is also an adaptive filter, as we will see that $\gamma_{k_1,k_2}(t)$ depends of the observation $\mathcal{S}x(t,f)$. Note that it can be shown that this estimator is the posterior mean estimator in the Bayesian setting.

We use in the following the present notation : $y \sim \mathcal{N}(0,\sigma^2)$ means y has a gaussian centered distribution of variance σ^2.

For a given frame index t and a given component couple $(k_1,k_2) = (q_1(t), q_2(t))$, the prior densities of both sources are gaussian, centered:

$$\mathcal{S}s_1(t,f) \sim \mathcal{N}(0, \mathrm{diag}(\sigma^2_{k_1}(f)))$$
$$\mathcal{S}s_2(t,f) \sim \mathcal{N}(0, \mathrm{diag}(\sigma^2_{k_2}(f)))$$

The observed process $\mathcal{S}x(t,f) = \mathcal{S}s_1(t,f) + \mathcal{S}s_2(t,f)$ is centered gaussian distributed with covariance matrix $\mathrm{diag}(\sigma^2_{k_1}(f) + \sigma^2_{k_2}(f))$.

$$\mathcal{S}x(t,f) \sim \mathcal{N}(0, \mathrm{diag}(\sigma^2_{k_1}(f)) + \sigma^2_{k_2}(f)))$$

As we have $\gamma_{k_1,k_2}(t) \propto p(\mathcal{S}x(t,f)|k_1,k_2)\cdot p(k_1)\cdot p(k_2)$, we get in the GMM case :

$$\gamma_{k_1,k_2}(t) \propto \omega_{k_1}\omega_{k_2}p_G(\mathcal{S}x(t,f); \{\sigma^2_{k_1}(f) + \sigma^2_{k_2}(f)\}) \quad (1)$$

In expression (1), we get an adaptive Wiener filtering scheme, as the weighting probability $\gamma_{k_1,k_2}(t)$ depends on the observed mixture $\mathcal{S}x(t,f)$, for a given t.

This yields the following estimation algorithm

Algorithm 1

For all frame indexes t
1: Compute for all couple of components (k_1,k_2) the posterior probability $\gamma_{k_1,k_2}(t) = p_t((k_1,k_2)|\mathcal{S}x(t_1,f),\ldots,\mathcal{S}x(t_N,f))$
2: Filter

$$\widehat{\mathcal{S}s}_1(t,f) = \left[\sum_{k_1,k_2}\gamma_{k_1,k_2}(t)\frac{\sigma^2_{k_1}(f)}{\sigma^2_{k_1}(f)+\sigma^2_{k_2}(f)}\right]\mathcal{S}x(t,f)$$
$$\widehat{\mathcal{S}s}_2(t,f) = \left[\sum_{k_1,k_2}\gamma_{k_1,k_2}(t)\frac{\sigma^2_{k_2}(f)}{\sigma^2_{k_1}(f)+\sigma^2_{k_2}(f)}\right]\mathcal{S}x(t,f)$$

Note that the couple of components (k_1,k_2) of the source model can also be seen as the component of the observed process model. As $\mathcal{S}x(t,f) = \mathcal{S}s_1(t,f) + \mathcal{S}s_2(t,f)$, the density of the observation is the convolution of both prior models. In the GMM context, this density is still a mixture of gaussian model with $Q = Q1 \times Q2$ components :

$$p(\mathcal{S}x(t,f)) = \sum_{k_1=1}^{Q_1}\sum_{k_2=1}^{Q_2}\omega_{k_1}\omega_{k_2}p_G(\mathcal{S}x(t,f); \{\sigma^2_{k_1}(f) + \sigma^2_{k_2}(f)\})$$

We will see now how the parameters $\{\sigma^2_{k_i}\}$ and $\{\omega_{k_i}\}$ are estimated in a training phase, in which excerpts of each source are provided.

3. TRAINING PHASE

In a first phase, we assume that we have at our disposal audio samples, which are representative of each source, in order to estimate the model parameters, that is the covariance matrix and the prior weight of each component.

In order to estimate these parameters we make use of the classical Expectation-Maximization (EM) algorithm [7].

In the following, we will not estimate directly the covariance matrices (PSD) as the variances of the observed process $\mathcal{S}s_i(t,f)$. We rather use the EM algorithm on the log module spectra $\log|\mathcal{S}s_i(t,f)|$ for a better accuracy in the segmentation of the data, that is the estimation of the posterior probability $\gamma_{k_i}(t)$ of observing component k_i of the source at time t. Then we have

$$\sigma^2_{k_i}(f) = \frac{\sum_t \gamma_{k_i}(t)|\mathcal{S}s_i(t,f)|^2}{\sum_t \gamma_{k_i}(t)}$$

The parameters estimated with the EM algorithm are the covariance matrices $\{\sigma^2_{k_i}(f)\}$ and the weights $\{\omega_{k_i}\}$, whereas the number of gaussian densities Q_i is set a priori.

We have now a complete framework for audio source separation using one single sensor. Let us see the changes that occur if we use hidden Markov Models instead of GMM.

4. HIDDEN MARKOV MODELS

In the case of gaussian mixture models, the prior weights of the gaussian densities are kept constant. Hidden Markov Models (HMM) with mixture of gaussian conditional densities, of order L, can be seen as a generalization of GMM, in which the prior weights at time t depend on the active HMM state, which corresponds to the component index $q_i(\tau)$, at previous times $\tau = t-1,\ldots,t-L$. The weights parameters are modeled in a matrix $\omega_{q_i(t),q_i(t-1),\ldots,q_i(t-L)}$.

Therefore the HMM density for the source s_i is:

$$p_i(\mathcal{S}s_i(t,f)) = \sum_{k_i=1}^{Q_i} \left[\omega_{k_i,q_i(t-1),\ldots,q_i(t-L)} \cdot p_G(\mathcal{S}s_i(t,f);\{\sigma^2_{k_i}(f)\})\right]$$

As we must compute the couple of components (k_1,k_2) posterior probability $\gamma_{k_1,k_2}(t)$, for each frame index t, it seems natural to use the forward backward algorithm [8] in the HMM models. This is the only difference between GMM and HMM modeling: once we have computed the probabilities $\gamma_{k_1,k_2}(t)$, we are able to compute the weighted Wiener-type filter.

The HMM models permit to take into account the a priori time dependencies between the modeled PSDs, through the state dependency structure.

We will compare in a real audio experiment the performance of GMM and HMM models. Note that we will consider only first order HMM model.

5. POST-PROCESSING

We propose now a post-processing step in order to improve the separation performances. At the end of the separation step, we have at our disposal two signals $\hat{s}_1$ and $\hat{s}_2$ which are supposed to be more separated than in the original mixture. Thus in a classical source separation point of view, we are now in a determined setting: we have two input signals for a post-processing source separation step.

To keep things simple, we have only used a decorrelation algorithm on the estimated sources, though we could have used standard Independent Component Analysis (ICA) techniques.

If we note $\hat{s} = [\hat{s}_1\hat{s}_2]$, $C = \frac{1}{T}\cdot\hat{s}^T\hat{s}$ is the 2×2 covariance matrix of the estimated sources. Then we set:

$$s^\star = \hat{s}\cdot C^{-1/2}$$

$s^\star$ represents the estimated sources after decorrelation, i.e. $s^{\star T}s^\star$ is diagonal.

Let us define now cirteria for the evaluation of the experiments on real audio sources.

6. EVALUATION CRITERIA

In the evaluation of the separation experiments, we need to define some criteria, in order to compare the performance of GMM and HMM models in various settings (different number of components for the model of each source). We will suppose that the two original sources s_1 and s_2 are uncorrelated and we denote their estimates $\hat{s}_1$ and $\hat{s}_2$.

Let us consider the orthogonal projection of the estimated sources over the vector space spanned by the real sources.

We may write $\hat{s}_1 = \alpha_1 s_1 + \alpha_2 s_2 + n_1$ and $\hat{s}_2 = \beta_1 s_1 + \beta_2 s_2 + n_2$.

We define a Source to Interference Ratio (SIR) as the ratio in dB between the source component $\alpha_1 s_1$ (in the case of the first source $\hat{s}_1$) and the interference component $\alpha_2 s_2$.

We also define a Source to Artefact Ratio (SAR) as the ratio between the sources components $\alpha_1 s_1 + \alpha_2 s_2$ and the noise component n_1. Note that these two components are orthogonal.

$$\mathrm{SIR}_1 = 20\log_{10}\left|\frac{\alpha_1}{\alpha_2}\right|\frac{\|s_1\|}{\|s_2\|} \quad \mathrm{SAR}_1 = 20\log_{10}\frac{\|\hat{s}_1 - n_1\|}{\|n_1\|}$$

$$\mathrm{SIR}_2 = 20\log_{10}\left|\frac{\beta_2}{\beta_1}\right|\frac{\|s_2\|}{\|s_1\|} \quad \mathrm{SAR}_2 = 20\log_{10}\frac{\|\hat{s}_2 - n_2\|}{\|n_2\|}$$

The SIR is a way to measure the residual of the other source in the estimation of each source, whereas the SAR is an estimate of the amount of distortion in each estimated signal. One may find more details about these measurements in [9].

7. EXPERIMENTAL ISSUES

We now discuss issues concerning the use of the log module spectra of the observed signal in the estimation of the weighting function $\gamma_{k_1,k_2}(t)$, in the separation phase. In section 3, for the training phase, we compute the posterior probabilities $\gamma_{k_i}(t)$ on the log module spectra $\log|\mathcal{S}s_i(t,f)|$ rather than on the signal, for accuracy reasons.

As a result, we should do the same thing in the separation phase for computing the weighting probabilities $\gamma_{k_1,k_2}(t)$ on the log module spectra of the observed process $\log|\mathcal{S}x(t,f)|$.

Therefore, we have studied experimentally the case of a complex random variable y with a centered gaussian distribution $y \sim \mathcal{N}(0,\sigma)$, in order to derive a gaussian model for the logarithm of the module of the variable y.

We denote $m(\sigma) = E(\log|y||\sigma)$ for the mean of the log module of a gaussian random variable. We have observed that

$$m(\sigma) = a_0 + \log(\sigma) \tag{2}$$

where $a_0 \approx -0.634$ with Monte-Carlo estimation.

Let us denote now $m_{k_i}(f)$ and $\beta_{k_i}(f)$ the means and the variance of the component k_i of the GMM/HMM model of the source s_i, which are estimated in the training phase. The mean $m_{k_1,k_2}(f)$ of the component (k_1,k_2) in the GMM model of the resulting mixture signal is according to formula 2 :

$$m_{k_1,k_2}(f) = \tfrac{1}{2}\log\left[\exp[2m_{k_1}(f)] + \exp[2m_{k_2}(f)]\right] \tag{3}$$
$$\beta_{k_1,k_2}(f) = \beta_{k_1}(f) + \beta_{k_2}(f)$$

where $\beta_{k_1,k_2}(f)$ is the variance of the component (k_1,k_2) in the GMM/HMM model of the composite signal. The definition 3 is consistent with the approximate formula used in [5] : $m_{k_1,k_2}(f) = \max\left[m_{k_1}(f), m_{k_2}(f)\right]$.

We use the following mixture model in the forward-backward recursion for the computation of the probabilities γ_{k_1,k_2}(t) :

$$p(\mathcal{S}x(t,f)|k_1,k_2) \approx$$
$$\sum_{k_1=1}^{Q_1}\sum_{k_2=1}^{Q_2}\omega_{k_1}\omega_{k_2}p_G\left(\log|\mathcal{S}x(t,f)|, m_{k_1,k_2}(f), \beta_{k_1,k_2}(f)\right)$$

8. EXPERIMENTAL STUDY

In the experimental setting, we take audio recordings from a jazz standart. The first source consists in the piano and bass part, whereas the second source consists of the drum part.

We use one minute of both excerpts as training data, that is for the estimation of the models parameters. The next 15 seconds of both sources are added to form a 'hand-made' source mixture. We estimate the sources in the separation phase from this audio mixture.

The excerpts are sampled at sampling rate of 11kHz. As an input of the STFT, We use a windowed signal frame of length 47 ms.

Note that the sources are approximately decorrelated, i.e $\frac{|(s_1,s_2)|}{\|s_1\|\|s_2\|} \approx 0.006$.

8.1. Evaluation

We evaluate the source to interference ratio (SIR) and the source to artecfact ratio (SAR) with varying numbers of components Q_1, Q_2 in the mixture models. We evaluate the GMM models and HMM models of order 1.

The scores with/without post-processing are given in table 1 for the SIR and table 2 for the SAR. Note that we have also given the SIR and SAR for the standard Wiener filtering, these tables, as this technique can be seen as a particular case of the proposed method with one unique mixture component per model.

state	source	GMM	HMM	GMM	HMM
		no post-processing		post-processing	
Wiener	piano	8.7	-	14.0	-
Wiener	drums	6.7	-	12.8	-
4	piano	10.5	10.5	36.4	35.7
4	drums	9.7	9.7	10.4	10.4
8	piano	11.0	10.8	31.5	29.3
8	drums	11.3	11.3	12.1	12.1
16	piano	11.8	11.6	25.1	23.0
16	drums	11.9	11.8	12.5	12.4

TAB. 1 –. SIR for each of the sources as a function of the number of components in each source model.

8.2. Discussion

As we compare the SIR and SAR (Signal to Interference Ratio and Signal to Artefact Ratio) with the GMM or HMM methods and with the standard Wiener filtering, we see that we obtain better results with the mixture models, according to the defined criteria.

We remark that, in the experiments, the SAR values are smaller than the SIR values in general. This is perhaps due

		no post-processing		post-processing	
state	source	GMM	HMM	GMM	HMM
Wiener	piano	7.8	-	4.7	-
Wiener	drums	5.8	-	2.6	-
4	piano	8.4	8.4	4.2	4.2
4	drums	5.9	5.8	5.5	5.5
8	piano	8.9	8.9	5.1	5.1
8	drums	5.9	5.8	5.7	5.5
16	piano	8.1	7.9	5.1	5.0
16	drums	5.4	5.1	5.2	4.9

TAB. 2 –. SAR for each of the sources as a function of the number of components in each source model.

to the stochastic aspect of the model, in which it is assumed that the phases of the STFT of the sources are random. As a consequence, the phase are not estimated and both estimated sources phases are equal to the mixture phase, at each frequency.

The post-processing step improves clearly the SIR, wheras it degrades slightly the SAR, in general.

We may also remark that the HMM models do not significantly improve the separation results over the GMM models, according to the SIR/SAR criteria.

9. CONCLUSION

We have presented and evaluated a statistical method for the separation of two audio sources using one single sensor. This method relies on a smooth adaptive Wiener filtering scheme and we make use of gaussian mixture models for the sources priors. The parameters of the a priori models are estimated in first phase and we have discussed the issues of this phase. Finally, the reported experiments show a significant improvement over the standard Wiener filter. Next step will be to work on a phase model for the sources and to introduce a psycho-acoustic model in the separation phase and in the evaluation criteria.

10. REFERENCES

[1] Aapo Hyvärinen, Patrik Hoyer, and Erkki Oja, "Image denoising by sparse code shrinkage," .

[2] A. Bijaoui, "Wavelets, gaussian mixtures and wiener filtering," *Signal Processing*, vol. 82, pp. 709–712, 2002.

[3] L. Benaroya, R. Gribonval, and F. Bimbot, "Non negative sparse representation for wiener based source separation with a single sensor," in *ICASSP*, 2003 (submitted).

[4] J. Portilla, V. Strela, M.J. WainWright, and E. Simoncelli, "Adaptive wiener denoising using a gaussian scale of mixture model in the wavelet domain," in *Proc. of the 8th international onference on Image Processing*, Thessaloniki,Greece, October 2001.

[5] Sam T. Roweis, "One microphone source separation," in *NIPS*, 2000, pp. 793–799.

[6] N. Wiener, *Extrapolation, interpolation and smoothing of stationary time series*, MIT press, 1949.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, 1977.

[8] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, 1989, vol. 77, pp. 257–285.

[9] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, "Proposals for performance measurement in source separation," in *ICA (submitted)*, 2003.