Audio source separation with a single sensor

Laurent BENAROYA, Frédéric BIMBOT, Rémi GRIBONVAL
IRISA (CNRS & INRIA), METISS, Campus de Beaulieu
35042 Rennes Cedex, France
E-mail : {lbenaroy,bimbot,remi}@irisa.fr

Abstract— **In this paper, we address the problem of audio source separation with one single sensor, using a statistical model of the sources. The approach is based on a learning step from samples of each source separately, during which we train gaussian scaled mixture models (GSMM). During the separation step, we derive Maximum A Posteriori (MAP) and/or Posterior Mean (PM) estimates of the sources, given the observed audio mixture (Bayesian framework). From the experimental point of view, we test and evaluate the method on real audio examples.**

Index Terms— **Gaussian mixture models, Wiener filtering, Audio source separation, Bayesian formalism**

I. Introduction

Source separation is an increasingly popular theme in the field of signal processing, especially since new tools, such as Independent Component Analysis (ICA) have been proposed, developed and improved [2], [7], [13], [5].

Independent Component Analysis has many applications, on biomedical, functional magnetic resonance imaging data for instance, as well as applications in speech processing and audio source separation.

The source separation problem can be formulated as an equation :

$$x_i = \sum_{j=1}^{m} a_{ij} s_j \tag{1}$$

where m sources s_j with amplitude factors a_{ij} are assumed to be summed to form a collection of n sensor signals x_i. This case is classically refered to as the *linear instantaneous mixture*. Note that hypotheses such as independence or non-gaussianity of the sources usually lead to a solution [11].

Two different cases may be distinguished :

1) The number of sensors n is greater or equal to the number of sources m. In this particular case, the estimation of the mixing matrix $A = (a_{ij})$ happens to be very useful, as the sources may be recovered via the pseudo-inverse of this matrix.
2) The number of sensors n is less than the number of sources m. In this case (known as the *under-determined* case), the estimation of the matrix A is not sufficient to recover the sources.

A. Presentation

The present article addresses an extreme situation of the second case (*under-determined*) [13]. We study here the case of a single sensor, with two sources, which is a very specific case, as the mixing equation is reduced to $x = s_1 + s_2$

Here are the main features of this work :

- We use a source model. Building a good model of each source is crucial and it must exploit some knowledge on the sources. In this respect, the approach may not be qualified as “blind” estimation, contrary to classical (even under-determined) cases. In this paper, we address the case of audio sources, of which we build (or assume) statistical models.
- There is a natural formalism for the single sensor case : the Bayesian formalism. This formalism is based on a statistical framework, as the phenomena we observe are variable. It makes it possible to take into account both the additive setting, which yields a likelihood function, and the source models, which provide *a priori* densities and correspond to prior knowledge on the problem. In practice, we consider a training step in which model parameters of each sources are estimated separately. We then make use of this prior information in the separation step.

Even though we consider, in this study, the special case of two sources with one single sensor, many results can be generalized to more sources (at least theoretically).

B. Formalism

In a probabilistic formalism, the sources can be estimated through a Maximum Likelihood (ML) estimate as the mixing equation (1) leads to the definition of a likelihood function :

$$(\widehat{s_1}, \widehat{s_2}) = \arg\max_{s_1, s_2} p(x|s_1, s_2) \tag{2}$$

where x is the observed signal, whereas s_i, $i = 1, 2$ are the sources which are to be estimated. The problem with the ML approach is that there are multiple solutions, since the system is underdetermined.

It is therefore natural to introduce the *a posteriori* probability distribution for the sources, in a Bayesian formalism :

$$p(s_1, s_2|x) \propto p(x|s_1, s_2) p_1(s_1) p_2(s_2), \tag{3}$$

where $p(x|s_1, s_2)$ is the likelihood function and $p_1(s_1), p_2(s_2)$ correspond to the prior knowledge about the sources. Here the sources are supposed to be independent, i.e. $p(s_1, s_2) = p_1(s_1) \cdot p_2(s_2)$.

Then, the Maximum A Posteriori estimator may lead to a solution for the source separation problem :

$$(\widehat{s_1}, \widehat{s_2}) = \arg\max_{s_1,\, s_2} p(s_1, s_2|x)$$

Relation (3) is the basis for the estimation of the sources, as it permits to take into account both the additive setting through the likelihood function, and the prior information about the sources via the *a priori* densities. The parameters of these prior densities (covariance matrices, for instance) are estimated in a off-line, training step.

Some former attempts have been made to solve the source separation problem with one single microphone [4]. In particular, the method proposed in [15] is close to our approach as it uses Hidden Markov Models and filter theory. Our work provides mathematically grounded algorithms and generalizes the approach to a wide range of statistical models and estimation criteria.

C. Bayesian approach

Several methods in ICA or even in "noisy" Principal Component Analysis (PCA) [16], [5], rely on the Bayesian formalism. In the case of the instantaneous linear mixture of m sources s into n sensors x, the basic equation (1) becomes : $x = As + b$, where b is some white noise (gaussian distributed for instance).

In this case, the noise distribution corresponds to the likelihood function, because we have : $p(x|s, A) = p(x - As) = p(b)$. In the particular case of Laplacian distributed sources (as prior distributions), the mixing matrix A may be estimated via the *Maximum A Posteriori* of the distribution of s conditionally to x (MAP criterion) :

$$\hat{s} = \arg\max_{s} p(s|x, A) = \arg\max_{s} p(x|s, A) \cdot p(s)$$

Generally speaking, when the prior laws are unknown, but the independence of the sources is assumed, the sources may be estimated through a semi-parametric approach [1].

In this study, the models behind the prior densities p_1, p_2 are more specific, though the formalism (i.e. the Bayesian point of view) is the same. In our approach, we use prior information about characteristic Power Spectral Densities (PSD) of each source in order to achieve the source separation. This information may be obtained in a prior training step on separated excerpts of the sources.

D. Organization of the paper

This article is organized as follows :

In section 2, we recall some basics of the Bayesian theory and we describe the classical Wiener filtering approach for stationary sources.

In section 3, we make use of the Bayesian formalism in order to derive Wiener estimators in the case of non-Gaussian priors.

In section 4, we present the resulting separation algorithm in the Short Term Fourier Transform (STFT) domain.

In section 5, we describe evaluation criteria which we use in the experiments.

Finally, in section 6, we test and evaluate the proposed approach on a real audio excerpt of Jazz music.

II. Wiener filtering

A. The Bayesian formalism

1) Framework: As explained in the introduction, the Bayesian formalism offers a natural framework in order to incorporate prior knowledge in an estimation problem. In this section, we recall how this framework can be used for estimating a parameter θ, given observed data x.

First, we assume that we are given a parametric statistical model, $f(x|\theta)$, where x represents the observed data. θ is the only unknown parameter (or set of parameters) which belongs to a *finite* dimensional vector space. The density $f(x|\theta)$, from which the data x is drawn, is called the likelihood function as a function of θ.

Then, we define the *a priori* distribution $\pi(\theta)$ of the parameter θ, which represents the knowledge we have about this parameter, before observing the data x. This leads to the definition of the *a posteriori* density, according to Bayes law :

$$p(\theta|x) \propto f(x|\theta)\pi(\theta)$$

From this distribution, the estimation of parameter θ is possible and, in a sense, the notion of a posteriori law is a key notion in the Bayesian theory.

2) Estimation and cost function: We study now the estimation of the parameter θ, according to the observed data x. To do this, we define a cost function $C(\alpha, \theta)$.

This cost $C(\alpha, \theta)$ represents the cost of replacing the true value of the parameter θ with its estimate α.

The estimation of the parameter θ is done by minimizing the mean cost over all possible values of θ, according to its posterior density.

$$\alpha_{opt} = \arg\min_{\alpha} \int_{\theta} C(\alpha, \theta) f(x|\theta)\pi(\theta) d\theta$$

In the case of a quadratic cost $C(\alpha, \theta) = \|\alpha - \theta\|_2^2$, the Bayesian estimator is the conditional Posterior Mean (PM) : $E(\theta|x)$. There exists another standard cost function $C(\alpha, \theta) = 1 - \delta_{\alpha-\theta}$ (δ is the Dirac distribution). In this case the corresponding Bayesian optimal estimator is the Maximum A Posteriori estimator (MAP) :

$$\alpha_{opt} = \arg\max_{\theta} f(x|\theta)\pi(\theta)$$

B. Bayesian formulation of the Wiener filter

Suppose s_1 and s_2 are two Gaussian processes, independent, centered and with covariance matrices Σ_1 and Σ_2. We observe a noisy realization of the sum of the two processes, $x = s_1 + s_2 + b$, where b is some Gaussian white noise of variance σ^2.

As presented in the introduction, we have the following likelihood function : $p(x|s_1, s_2) = p(b)$ and prior density : $p(s_1, s_2) = p(s_1) \cdot p(s_2)$. If we further suppose that the noise component is Gaussian distributed, the likelihood function $p(x|s_1, s_2)$ becomes :

$$p(x|s_1, s_2) = p(b) = \mathrm{g}(x - s_1 - s_2, \sigma^2 I),$$

where $\mathrm{g}(y, \Sigma)$ is the gaussian centered distribution :

$$\mathrm{g}(y, \Sigma) = \frac{1}{(2\pi)^{N/2}} \frac{1}{\sqrt{|\det \Sigma|}} \exp\left[-\frac{1}{2} y^T \Sigma^{-1} y\right]$$

N being the dimension of the observation y.

Concerning the prior densities, we may assume that $p(s_1) = \mathrm{g}(s_1, \Sigma_1)$, $p(s_2) = \mathrm{g}(s_2, \Sigma_2)$. In this setting, the likelihood $p(x|s_1, s_2)$ is the parametric law of the observation x, whereas s_1 and s_2 are the parameters to be estimated. $p(s_1)$ and $p(s_2)$ are the a priori laws over the parameters, which represents knowledge about these parameters before observing x. In this section, we assume a Gaussian a priori law.

Relying on Bayes law, the following expression for the a posteriori law can be derived :

$$\begin{aligned} -\log p(s_1, s_2|x) &= \tfrac{1}{2\sigma^2} \|x - s_1 - s_2\|_2^2 \\ &+ \tfrac{1}{2} s_1^T \Sigma_1^{-1} s_1 + \tfrac{1}{2} s_2^T \Sigma_2^{-1} s_2 + cte \end{aligned} \tag{4}$$

We deduce the MAP estimator for s_1 and s_2 from this formula :

$$\begin{aligned} \widehat{s}_1 &= \Sigma_1 [\Sigma_1 + \Sigma_2 + \sigma^2 I]^{-1} x \\ \widehat{s}_2 &= \Sigma_2 [\Sigma_1 + \Sigma_2 + \sigma^2 I]^{-1} x \end{aligned}$$

In the case of a "vanishing" noise, i.e. $\sigma \to 0$, the estimator converges towards the Wiener estimator.

From expression (4), we see that the posterior distribution $p(s_1, s_2|x)$ is a gaussian distribution, as the expression inside the brackets is a quadratic form in s_1 and s_2. We conclude that the MAP and PM estimators are, in that case, identical.

C. Stationary Processes

In the specific case when s_1 and s_2 are stationary and (approximately) circular processes (i.e. with a Toeplitz covariance matrix) s_1 and s_2, the basis $\mathcal{B}$ which makes both covariance matrices diagonal is the discrete Fourier basis, which vectors are $\left\{\mathcal{F}f[n] = \frac{1}{\sqrt{N}} \exp\left(\frac{2\pi i f n}{N}\right)\right\}_{0 \le f < N}$, where $\mathcal{F}$ denotes the Discrete Fourier Transform operator and f denotes the frequency index.

In this case, the Wiener filtering can be interpreted as the following operation in the frequency domain :

$$\begin{aligned} \widehat{\mathcal{F}s_1}[f] &= \frac{\sigma_1^2[f]}{\sigma_1^2[f] + \sigma_2^2[f]} \cdot \mathcal{F}x[f] \\ \widehat{\mathcal{F}s_2}[f] &= \frac{\sigma_2^2[f]}{\sigma_1^2[f] + \sigma_2^2[f]} \cdot \mathcal{F}x[f] \end{aligned}$$

D. Limits and extensions

Let us recall the set of hypotheses made so far :

- The a priori knowledge concerning the sources is reduced to the knowledge of the covariance matrices, which corresponds to Power Spectral Densities (PSD), in the stationary case.
- The stochastic processes s_1 and s_2 are assumed to be Gaussian; equivalently we restrict the problem to linear estimators.
- Both processes s_1 and s_2 are stationary and circular.

As audio signals are generally non-Gaussian and non-stationary, the previous method may not be applied directly. The approach must be generalized to other prior densities, through the Bayesian framework.

This suggests to extend classical Wiener filtering to different kind of prior densities, in particular to non-Gaussian unimodal densities, to Gaussian mixture models and even to more complex models.

III. Extensions of Wiener filters to non-Gaussian priors

A. Non-Gaussian unimodal densities

The Wiener filter approach can be extended to other families of unimodal densities, for instance generalized gaussian densities :

$$\mathrm{G}(y, \alpha, \Sigma) \propto \exp\left[-\beta(\alpha) \left\|\Sigma^{-1/2} y\right\|_\alpha^\alpha\right]$$

where $\|y\|_\alpha^\alpha = \sum_k |y[k]|^\alpha$. We recall that $\beta(\alpha) = \left[\frac{\Gamma(3/\alpha)}{\Gamma(1/\alpha)}\right]^{\alpha/2}$

The Bayesian model now takes the following form :

$$\begin{array}{ll} p(x|s_1, s_2) = \mathrm{g}(x - s_1 - s_2, \sigma^2 I) & \text{likelihood function} \\ \left.\begin{array}{rcl} p(s_1) &=& \mathrm{G}(s_1, \alpha_1, \Sigma_1) \\ p(s_2) &=& \mathrm{G}(s_2, \alpha_2, \Sigma_2) \end{array}\right\} & \text{prior densities} \end{array}$$

The a priori law of the sources s_1 and s_2 are thus generalized gaussian densities.

Using Bayes law, the a posteriori law becomes :

$$\begin{aligned} -\log p(s_1, s_2|x) &= \frac{1}{2\sigma^2} \|x - s_1 - s_2\|_2^2 \\ + \beta(\alpha_1) \left\|\Sigma_1^{-1/2} s_1\right\|_{\alpha_1}^{\alpha_1} &+ \beta(\alpha_2) \left\|\Sigma_2^{-1/2} s_2\right\|_{\alpha_2}^{\alpha_2} + cte \end{aligned}$$

It is sometimes possible to find an expression (in some cases, an analytic one) for the MAP and PM estimators of s_1 and s_2. Let us have a look at the MAP estimator in some particular cases.

1) Particular cases:

a) Both sources have laplacian prior densities: In the case $\alpha_1 = \alpha_2 = 1$, i.e. both prior densities are laplacian laws and the covariance matrices are diagonal, we obtain the following MAP estimators, in the noiseless case :

$$\begin{aligned} \hat{s}_1[k] &= \left\{\begin{array}{ccc} x[k] & \text{if} & \sigma_1[k] > \sigma_2[k] \\ 0 & \text{otherwise} & \end{array}\right. \\ \hat{s}_2[k] &= \left\{\begin{array}{ccc} x[k] & \text{if} & \sigma_2[k] > \sigma_1[k] \\ 0 & \text{otherwise} & \end{array}\right. \end{aligned}$$

b) One laplace source and one gaussian source: Assuming now that source s_1 has a laplacian prior density, i.e. $\alpha_1 = 1$, with diagonal covariance matrix, whereas s_2 is a gaussian white noise of variance σ_2^2, that is $\alpha_2 = 2$.

Then the MAP estimator for s_1 is the coefficients shrinkage proposed by D. Donoho in [8] :

$$\hat{s}_1[k] = \operatorname{sign}(x[k]) \cdot \max[|x[k]| - \lambda[k], 0], \tag{5}$$

where $\lambda[k] = \frac{\sqrt{2}\sigma_2^2}{\sigma_1[k]}$.

In this case, the second source may be considered as a noise and the expression (5) may be interpreted as a reduction of the corrupted observed signal x from a quantity proportional to the noise variance. If the sources are expressed in a wavelet basis, this is often refered to as wavelet-shrinkage and this is a powerful tool for denoising purposes.

c) A more general case : $\alpha_1 = \alpha_2 = \alpha$: In this case (same generalized gaussian density for each source), it is possible to define a function $h(r;\alpha)$ if both covariance matrix are diagonal. $h(r;\alpha)$ is the function :

$$h(r;\alpha) = \frac{r^{\left(\frac{\alpha}{\alpha-1}\right)}}{1 + r^{\left(\frac{\alpha}{\alpha-1}\right)}},$$

in the case $\alpha > 1$. We obtain (noiseless case) :

$$\begin{aligned} \widehat{s}_1[k] &= h\left(\frac{\sigma_1[k]}{\sigma_2[k]};\alpha\right) x[k] \\ \widehat{s}_2[k] &= \left[1 - h\left(\frac{\sigma_1[k]}{\sigma_2[k]};\alpha\right)\right] x[k] \end{aligned}$$

This is just a generalization of the Wiener filter formula with a different shape for the weighting function h.

B. The case of Gaussian Mixture Models

The above developments can be viewed as examples of what can be done using the Bayesian framework, in the case of unimodal densities.

For dealing with non-stationary signals, it is necessary to consider other families of models for the sources. In this section, we study the case of Gaussian Mixture prior densities (GMM priors) [6], in line with former work in the field of speech processing, where parent approaches have been used to enhance the robustness of speech recognition in noisy environments (see for instance : [17], [18], [19]).

$$\mathcal{G}(y, \{\varpi^{(i)}\}, \{\Sigma^{(i)}\}) = \sum_{i=1}^{K} \varpi^{(i)} \mathrm{g}(y, \Sigma^{(i)}) \qquad (6)$$

where g is the gaussian function and $\sum_{i=1}^{K} \varpi^{(i)} = 1$.

As a generative model, the Gaussian Mixture model assumes that an observation is obtained by first selecting one active component within the K gaussians in the mixture (following the probability distribution $\{\varpi^{(i)}\}$) and then generating a gaussian observation following $\mathrm{g}(y, \Sigma^{(i)})$ for the active component.

For source separation, the Gaussian Mixture model permits to deal with multiple covariance matrices, that is multiple PSD (Power Spectral Densities) shapes, in the case of frequency domain filtering.

In the Bayesian formalism, we obtain the following prior densities :

$$\begin{aligned} p(s_1) &= \sum_{i=1}^{K_1} \varpi_1^{(i)} \frac{\exp\left[-\frac{1}{2} s_1^T {\Sigma_1^{(i)}}^{-1} s_1\right]}{(2\pi)^{N/2} |det(\Sigma_1^{(i)})|^{1/2}} \\ p(s_2) &= \sum_{j=1}^{K_2} \varpi_2^{(j)} \frac{\exp\left[-\frac{1}{2} s_2^T {\Sigma_2^{(j)}}^{-1} s_2\right]}{(2\pi)^{N/2} |det(\Sigma_2^{(j)})|^{1/2}} \end{aligned}$$

with $\sum_{i=1}^{K_1} \varpi_1^{(i)} = \sum_{j=1}^{K_2} \varpi_2^{(j)} = 1$.

Here, the MAP estimation is not tractable directly. In order to get back to the gaussian case (which is solved with Wiener filters), we introduce hidden variables q_1 and q_2 which are associated with the active components in both GMM models, i.e. the gaussian densities from which the sources data were most likely generated. This is a typical incomplete data setting.

In other words, the following likelihood and prior densities for the hidden q_i process ($i = 1, 2$) are considered :

$$\begin{aligned} p(s_i | q_i = k) &= \frac{\exp\left[-\frac{1}{2} s_i^T {\Sigma_i^{(k)}}^{-1} s_i\right]}{(2\pi)^{N/2} |det(\Sigma_i^{(k)})|^{1/2}} \\ p(q_i = k) &= \varpi_i^{(k)} \end{aligned}$$

The estimators are thus calculated conditionally to the hidden state couple (q_1, q_2).

1) First step : state estimation: As the couple of states (q_1, q_2) is generally unknown, we have to estimate this couple.

If the states are $q_1 = i$ and $q_2 = j$ (that is to say, if we know the active components in both mixture models), then s_1 has a gaussian distribution conditionally to q_1, of covariance matrix $\Sigma_1^{(i)}$ and s_2 has also a gaussian distribution conditionally to q_2 with covariance matrix $\Sigma_2^{(j)}$. We deduce that the sum $x = s_1 + s_2 + b$ has gaussian distribution conditionally to (q_1, q_2) with covariance matrix $\Sigma_1^{(i)} + \Sigma_2^{(j)} + \sigma^2 I$.

We deduce then the following posterior formula :

$$\begin{aligned} p(i, j | x) &\propto p(x | i, j) \cdot p(i) \cdot p(j) \\ &\propto \varpi_1^{(i)} \varpi_2^{(j)} \mathrm{gauss}(x, \Sigma_1^{(i)} + \Sigma_2^{(j)} + \sigma^2 I) \end{aligned}$$

This is the a posteriori law for the couple of components (i, j) for both mixture models, conditionally to the observed process x. We will note in the following $\gamma_{i,j}(x) = p(i, j | x)$, which is the a posteriori probability that the components (i, j) are active in each respective GMM, when observing x.

2) Second step : construction of the filters: If the active states q_1 and q_2 are known, then the problem can be solved by the Wiener filter approach, conditionally to the couple (q_1, q_2), as both priors are conditionally gaussian.

We have :

$$\begin{aligned} -2 \log p(s_1, s_2 | x, i, j) &= \frac{1}{\sigma^2} \|x - s_1 - s_2\|_2^2 \\ &+ s_1^T [\Sigma_1^{(i)}]^{-1} s_1 + s_2^T [\Sigma_2^{(j)}]^{-1} s_2 + cte \end{aligned}$$

If $q_1 = i$ and $q_2 = j$ are known, we have the conditional Bayesian (Wiener) estimator (as we have seen previously, the conditional MAP and PM estimators coincide) :

$$\begin{aligned} E(s_1 | i, j) &= \Sigma_1^{(i)} [\Sigma_1^{(i)} + \Sigma_2^{(j)} + \sigma^2 I]^{-1} x \\ E(s_2 | i, j) &= \Sigma_2^{(j)} [\Sigma_1^{(i)} + \Sigma_2^{(j)} + \sigma^2 I]^{-1} x \end{aligned}$$

a) Maximum A Posteriori estimation: When the active components $q_1 = i$, $q_2 = j$ are not known, they can be estimated as the MAP estimation of $\gamma_{i,j}(x)$ yielding one active component per GMM source model. In that case, we fall back on the Wiener filter setting, using the estimated couple of states. The approach can be understood as an *adaptive* Wiener filtering process.

b) Posterior Mean estimator: We may also estimate the sources s_1 and s_2 through the PM estimator [9].

As we have from Bayes law :

$$\begin{aligned} p(s_1|x) &\propto \sum_{i,j} p(s_1|x,i,j)p(q_1=i,q_2=j|x) \\ &\propto \sum_{i,j} p(s_1|x,i,j)\gamma_{i,j}(x) \end{aligned}$$

We deduce the following PM estimator :

$$\begin{aligned} E(s_1|x) &= \int_{s_1} s_1 p(s_1|x) ds_1 \\ &= \sum_{i,j} \gamma_{i,j}(x) \left[\int_{s_1} s_1 p(s_1|x,i,j) ds_1\right] \\ &= \sum_{i,j} \gamma_{i,j}(x)\, E(s_1|x,i,j) \end{aligned}$$

Finally :

$$E(s_1|x) = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2} \gamma_{i,j}(x)\cdot \Sigma_1^{(i)}[\Sigma_1^{(i)}+\Sigma_2^{(j)}+\sigma^2 I]^{-1}\cdot x$$

and similarily for s_2.

Moreover, relying on the above developments :

$$\gamma_{i,j}(x) \quad \propto \quad \varpi_1^{(i)}\varpi_2^{(j)}\mathrm{g}(x,\Sigma_1^{(i)}+\Sigma_2^{(j)}+\sigma^2 I),$$

with $\sum_{i,j}\gamma_{i,j}(x)=1$.

Thus, the first step consists in computing the posterior probabilities $\gamma_{i,j}(x)$, followed by the computation of weighted Wiener filters. The second step consists in filtering the sources with this adapted filter, with weight coefficients $\gamma_{i,j}(x)$ which thus depend on the observed process x.

3) HMM models: It must be noted that the generalized Wiener filter with GMM models can be extended to HMM models (Hidden Markov Models). Indeed, the only difference is that the weighting probabilities $\gamma_{i,j}(x)$ must then be computed through a forward-backward algorithm, which may result in a greater algorithmic cost[1].

4) Limitations of the GMM model: In the context of audio processing, we may observe the same sound corresponding to a similar PSD shape, repeated at different amplitudes and time indexes. If the GMM models are used as described above, there has to be as many gaussian components as there are different possible amplitudes, although they correspond to the same sound. This is quite restrictive.

This is why we have considered a more elaborate model : the Gaussian Scaled Mixture Model (GSMM), in order to separate the variance shape (PSD), and the amplitude information (gain factor).

C. Gaussian scaled mixture models

The Gaussian Scaled mixture model is a mixture of gaussian scaled densities [14].

A gaussian scaled density corresponds to a random variable of the form $g_a = \sqrt{a}\cdot g$, where g is a gaussian distributed vector variable with variance σ^2 and a is a non-negative scalar random variable, which may be drawn according to a prior density $p_0(a)$.

Thus the density of the gaussian scaled variable y is :

$$g_a(y|a) = \frac{1}{\sqrt{2\pi a}\sigma}\exp\left[-1/2\frac{\|y\|_2^2}{a\sigma^2}\right].$$

The marginal law is :

$$g_a(y) \propto \int_a \frac{1}{\sqrt{2\pi a}\sigma}\exp\left[-1/2\frac{\|y\|_2^2}{a\sigma^2}\right]p_0(a)da.$$

A gaussian scaled mixture model takes therefore the following form :

$$\begin{aligned} p(s_1|a_1^1,\ldots,a_{K_1}^1) &= \sum_{i=1}^{K_1}\varpi_1^{(i)}\frac{\exp\left[-\frac{1}{2}s_1^T\left(a_i^1\Sigma_1^{(i)}\right)^{-1}s_1\right]}{(2\pi a_i^1)^{N/2}|det(\Sigma_1^{(i)})|^{1/2}} \\ p(s_2|a_1^2,\ldots,a_{K_2}^2) &= \sum_{j=1}^{K_2}\varpi_2^{(j)}\frac{\exp\left[-\frac{1}{2}s_2^T\left(a_j^2\Sigma_2^{(j)}\right)^{-1}s_2\right]}{(2\pi a_j^2)^{N/2}|det(\Sigma_2^{(j)})|^{1/2}} \end{aligned}$$

Conditionally to $q_1=i$, $q_2=j$, a_i^1 and a_j^2, the Bayesian estimator (MAP or PM) is (cf. Wiener filter) :

$$E(s_1|x,i,j,a_i^1,a_j^2) = a_i^1\Sigma_1^{(i)}[a_i^1\Sigma_1^{(i)}+a_j^2\Sigma_2^{(j)}+\sigma^2 I]^{-1}x$$
$$E(s_2|x,i,j,a_i^1,a_j^2) = a_j^2\Sigma_2^{(j)}[a_i^1\Sigma_1^{(i)}+a_j^2\Sigma_2^{(j)}+\sigma^2 I]^{-1}x$$

Conditionally to $a_1^1,\ldots,a_{K_1}^1$ and $a_1^2,\ldots,a_{K_2}^2$, the weighting probabilities are :

$$\gamma_{i,j,a_i^1,a_j^2}(x) \propto \varpi_1^{(i)}\varpi_2^{(j)}\mathrm{gauss}(x,a_i^1\Sigma_1^{(i)}+a_j^2\Sigma_2^{(j)}+\sigma^2 I)$$

as $a_i^1\Sigma_1^{(i)}+a_j^2\Sigma_2^{(j)}+\sigma^2 I$ is the covariance matrix of the observed process, conditionally to the couple of states and the amplitudes.

For the posterior mean Bayesian estimator, we should integrate these estimates over all possible values of the amplitude parameters, that is :

$$\begin{aligned} E(s_1|x) = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\int_{a_i^1}\int_{a_j^2}\gamma_{i,j,a_i^1,a_j^2}(x)\times \\ E(s_1|x,i,j,a_i^1,a_j^2)p_0(a_i^1)p_0(a_j^2)da_i^1da_j^2 \end{aligned}$$

As the integrals may be untractable, we use a maximum likelihood estimation to determine the coefficients a_i^1 and a_j^2, under a positivity constraint and we set the amplitude coefficients to this value instead of integrating out.

We use the following estimation formula :

$$(\widehat{a_i^1},\ \widehat{a_j^2}) = \arg\max_{a_1\geq 0,\ a_2\geq 0}\gamma_{i,j,a_i^1,a_j^2}(x) \qquad (7)$$

which can be seen as a reweighted positive least square estimate.

[1] The algorithmic complexity of the algorithm with GMM models (which can be viewed as HMM models of order 0) is of order $O(Q_1\cdot Q_2)$, where Q_1 and Q_2 are the number of gaussian components in each source model. With fully-connected HMM models of order p, the complexity becomes $O(Q_1^{p+1}\cdot Q_2^{p+1})$. As a result, the algorithmic complexity with HMM models may be very high and even untractable in the case of HMM models of order greater than one, unless they are only sparsely connected.

IV. Separation algorithm

As we aim to separate audio sources, which are locally stationary in general, it is natural to work with the Short-Term Fourier Transform (STFT) denoted by $\mathcal{S}$. As this transform is linear, the additive setting remains : $\mathcal{S}x(t,f) = \mathcal{S}s_1(t,f) + \mathcal{S}s_2(t,f) + \mathcal{S}b(t,f)$. The covariance matrices $\Sigma_1^{(i)}$ and $\Sigma_2^{(j)}$ are assumed to be diagonal, with running element $\sigma_1^{(i)}(f)^2$ and $\sigma_2^{(j)}(f)^2$ respectively.

A. GMM models

We note $\gamma_{i,j}(x) = \gamma_{i,j}(t)$, the weighting probabilities corresponding to the observed frame $\mathcal{S}x(t,f)$ at time index t. The separation algorithm with the GMM models is given in the Algorithm 1.

Algorithm 1

```
Each source $s_1$ and $s_2$ is characterised by a Gaussian Mixture Model $\{\varpi_1^{(i)}, \sigma_1^{(i)}(f)\}_{1\le i\le K_1}$ and $\{\varpi_2^{(j)}, \sigma_2^{(j)}(f)\}_{1\le j\le K_2}$. The noise level $\sigma$ is set to an arbitrarily small value.
  1) For each frame index $t$, compute the weighting probabilities :
       $\gamma_{i,j}(t) \propto \varpi_1^{(i)}\varpi_2^{(j)} \times \prod_f \text{gauss}(|\mathcal{S}x(t,f)|, \sigma_1^{(i)}(f)^2 + \sigma_2^{(j)}(f)^2 + \sigma^2)$,
     with $\sum_{i,j}\gamma_{i,j}(t) = 1$.
  2) Then compute the posterior mean estimator :
       $\widehat{\mathcal{S}s_1}(t,f) = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\gamma_{i,j}(t)\frac{\sigma_1^{(i)}(f)^2}{\sigma_1^{(i)}(f)^2+\sigma_2^{(j)}(f)^2+\sigma^2}\mathcal{S}x(t,f)$,
       $\widehat{\mathcal{S}s_2}(t,f) = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\gamma_{i,j}(t)\frac{\sigma_2^{(j)}(f)^2}{\sigma_1^{(i)}(f)^2+\sigma_2^{(j)}(f)^2+\sigma^2}\mathcal{S}x(t,f)$.
```

B. GSMM models

In the STFT setting, conditionally to the pair of states (i,j) and to the amplitude parameters a_i^1, a_j^2, the sources are Gaussian centered processes. Therefore, the observed mixture is also a Gaussian centered process, of diagonal covariance $a_i^1\sigma_1^{(i)}(f)^2 + a_j^2\sigma_2^{(j)}(f)^2 + \sigma^2$. Then, we have the following likelihood function :

$$p(x|i,j,a_i^1,a_j^2) \approx \tag{8}$$

$$\frac{\exp\left[-\sum_f \frac{|\mathcal{S}x(t,f)|^2}{2(a_i^1\sigma_1^{(i)}(f)^2+a_j^2\sigma_2^{(j)}(f)^2+\sigma^2)}\right]}{\prod_f\sqrt{2\pi(a_i^1\sigma_1^{(i)}(f)^2+a_j^2\sigma_2^{(j)}(f)^2+\sigma^2)}}. \tag{9}$$

The amplitude coefficients can be computed in a Maximum Likelihood scheme, under positivity constraints. It can be shown [3] that equation (7) can be solved by finding (a_i^1, a_j^2) so as to solve the following system :

$$\sum_f a_i^1\sigma_1^{(i)}(f)^2\frac{|\mathcal{S}x(t,f)|^2-(a_i^1\sigma_1^{(i)}(f)^2+a_j^2\sigma_2^{(j)}(f)^2+\sigma^2)}{(a_i^1\sigma_1^{(i)}(f)^2+a_j^2\sigma_2^{(j)}(f)^2+\sigma^2)^2} = 0$$

$$\sum_f a_j^2\sigma_2^{(j)}(f)^2\frac{|\mathcal{S}x(t,f)|^2-(a_i^1\sigma_1^{(i)}(f)^2+a_j^2\sigma_2^{(j)}(f)^2+\sigma^2)}{(a_i^1\sigma_1^{(i)}(f)^2+a_j^2\sigma_2^{(j)}(f)^2+\sigma^2)^2} = 0$$

These equations are obtained by differentiating the logarithm of equation (8) with respect to the amplitude parameters, and introducing Lagrange multipliers in order to incorporate the positivity constraints. They can be solved through an iterative procedure, where the denominator is kept constant [12], leading to the first step as described in Algorithm 2.

The estimation of the amplitude parameters a_1 and a_2 can be interpreted as a match of the squared spectral module of the STFT process $\mathcal{S}x(t,f)$ with the estimated variances $a_i^1\sigma_1^{(i)}(f)^2 + a_j^2\sigma_2^{(j)}(f)^2 + \sigma^2$, under positivity constrains.

The separation algorithm with the GSMM models is summarized in the Algorithm 2.

Algorithm 2

```
Each source $s_1$ and $s_2$ is characterised by a Gaussian Mixture Model $\{\varpi_1^{(i)}, \sigma_1^{(i)}(f)\}_{1\le i\le K_1}$ and $\{\varpi_2^{(j)}, \sigma_2^{(j)}(f)\}_{1\le j\le K_2}$. The noise level $\sigma$ is set to an arbitrarily small value.
  1) For each frame index $t$, compute the amplitude coefficients, by the following iterative procedure :
     a) Initialize $[a_i^1(t)]_0$ and $[a_j^2(t)]_0$
     b) At each iteration $\ell$
        • evaluate $[r(t,f)]_\ell^2 = [a_i^1(t)]_{\ell-1}\ \sigma_1^{(i)}(f)^2 + [a_j^2(t)]_{\ell-1}\ \sigma_2^{(j)}(f)^2 + \sigma^2$
        • update $[a_i^1(t)]_\ell = [a_i^1(t)]_{\ell-1} \times \sum_f \frac{\sigma_1^{(i)}(f)^2}{[r(t,f)]_\ell^2}\frac{|\mathcal{S}x(t,f)|^2}{[r(t,f)]_\ell^2} / \sum_f \frac{\sigma_1^{(i)}(f)^2}{[r(t,f)]_\ell^2}$
        • update $[a_j^2(t)]_\ell = [a_j^2(t)]_{\ell-1} \times \sum_f \frac{\sigma_2^{(j)}(f)^2}{[r(t,f)]_\ell^2}\frac{|\mathcal{S}x(t,f)|^2}{[r(t,f)]_\ell^2} / \sum_f \frac{\sigma_2^{(j)}(f)^2}{[r(t,f)]_\ell^2}$
        • if *convergence* then $a_i^1(t) = [a_i^1(t)]_\ell$ and $a_j^2(t) = [a_j^2(t)]_\ell$ else *iterate* $(\ell \leftarrow \ell + 1)$
  2) Compute the weighting probabilities $\gamma_{i,j}(t)$ :
       $\gamma_{i,j}(t) \propto \varpi_1^{(i)}\varpi_2^{(j)} \times \prod_f \text{gauss}(|\mathcal{S}x(t,f)|, a_i^1\sigma_1^{(i)}(f)^2 + a_j^2\sigma_2^{(j)}(f)^2 + \sigma^2)$,
     with $\sum_{i,j}\gamma_{i,j}(t) = 1$.
  3) Finally, filter the observed frame :
       $\widehat{\mathcal{S}s_1}(t,f) = \sum_{i,j}\gamma_{i,j}(t)\frac{a_i^1\sigma_1^{(i)}(f)^2}{a_i^1\sigma_1^{(i)}(f)^2+a_j^2\sigma_2^{(j)}(f)^2+\sigma^2}\mathcal{S}x(t,f)$,
       $\widehat{\mathcal{S}s_2}(t,f) = \sum_{i,j}\gamma_{i,j}(t)\frac{a_j^2\sigma_2^{(j)}(f)^2}{a_i^1\sigma_1^{(i)}(f)^2+a_j^2\sigma_2^{(j)}(f)^2+\sigma^2}\mathcal{S}x(t,f)$.
```

V. Evaluation criteria

For the evaluation of the separation experiments, we need to define some criteria, in order to compare the performance of GMM models in various settings (different numbers of components for the model of each source). We 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 actual mixture $\alpha_1 s_1 + \alpha_2 s_2$ and the noise component n_1. Note that these two components are supposed to be orthogonal.

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

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

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 measures in [10].

VI. Experimental study

In the experimental setting, we work on two tracks of a jazz piece, provided separately on a CD designed to learn how to play jazz. A first track contains the piano and bass part, whereas the second track consists of the drum part. Both tracks are consistent with each other, i.e. when they are mixed, they form a coherent piece of music.

We use 45 seconds of each excerpt separately as training data, for estimating both source model parameters (PSD vectors and prior weights in the GMM model) : one model for the piano+bass track and another model for the drums. This is done using a conventional Expectation-Maximisation procedure for optimising the training data likelihood (Maximum Likelihood criterion).

The next 15 seconds of music are mixed by adding both tracks. This excerpt is different from the training excerpts. We estimate the sources in the separation step from the audio mixture, using as prior knowledge the source models estimated in the training phase.

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

Note that the sources are approximately decorrelated, as $10 \log_{10} \frac{|\langle s_1, s_2 \rangle|}{\|s_1\| \|s_2\|} = -16$ dB. Indeed, although belonging to the same piece, the sources do not show any short-term correlation, though they obviously are not completely independent.

A. Evaluation

We evaluate the Source to Interference Ratio (SIR) and the Source to Artefact Ratio (SAR) with various numbers of components Q_1, Q_2 in the mixture models. We evaluate the GMM models and Gaussian Scaled Mixture Models (GSMM).

The performances are reported in table I for the SIR and table II for the SAR. Note that we have also given the SIR and SAR for the standard Wiener filtering, in these tables, as this technique can be seen as a particular case of the proposed method with a single mixture component per model.

B. Discussion

As the number of gaussian components in each source model goes from 1 (Wiener standard setting) to 4 components and then 8 components, the SIR and SAR seem to improve.

criterion →		PM		MAP	
state	source	GMM	GSMM	GMM	GSMM
Wiener	piano	8.7	4.7	15.5	4.4
Wiener	drums	6.7	19.6	0.7	13.8
4	piano	10.5	11.0	20.0	10.5
4	drums	9.7	18.5	2.8	18.2
8	piano	11.0	11.1	20.0	10.7
8	drums	11.3	18.2	4.0	18.1
16	piano	11.8	12.9	16.9	12.6
16	drums	11.9	16.9	4.2	16.3

TABLE I

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

criterion →		PM		MAP	
state	source	GMM	GSMM	GMM	GSMM
Wiener	piano	7.8	12.3	1.8	12.4
Wiener	drums	5.8	0.0	14.7	-0.5
4	piano	8.4	8.7	2.0	8.6
4	drums	5.9	4.0	9.6	3.7
8	piano	8.9	8.5	3.4	8.7
8	drums	5.9	4.0	8.9	4.0
16	piano	8.1	8.6	3.5	8.4
16	drums	5.4	5.3	8.3	5.0

TABLE II

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

Then with 16 components, the SIR and SAR decrease in some cases (and for some particular estimators) or increase in other cases. This may be interpreted as a consequence of model overfitting, although it might come also from initialization problems in the training step (EM algorithm).

For the GSMM approach with 16 components for each source model, the SIR reaches approximately 12 dB for the piano+bass source and 16 dB for the drum source, with an SAR in the range of 9 and 5 dB respectively. These figures globally represent an improvement compared to the standard Wiener filtering technique, which shows an advantage in using source models that are able to track their statistical behaviours.

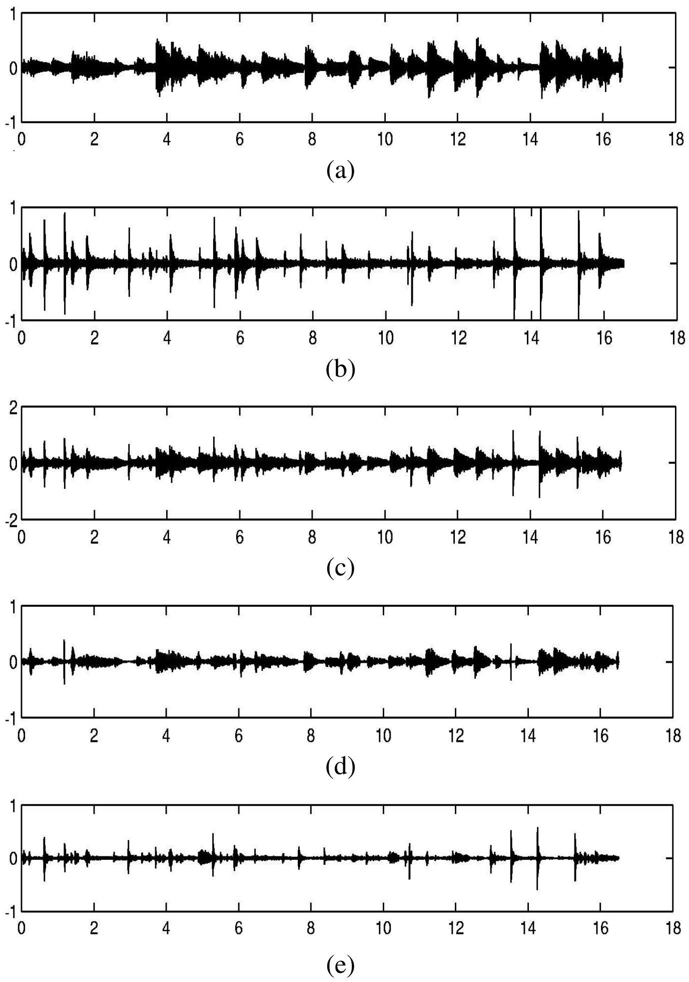
We may remark that in the GSMM case, the MAP criterion gives slighlty poorer results compared to the PM criterion, although it is computationally less expensive. In the GMM case, the MAP criterion gives poor results.

The GSMM model seems to improve the SIR results compared to the GMM model, in particular for the drum source, at the cost of a slight SAR decrease.

It must be underlined that the trends observed in our experiments are undoubtedly dependent on the statistical properties of the two sources used in this study. A more comprehensive experimental investigation, using various sources and different families of models will be necessary before drawing conclusions with a more general significance.

VII. Conclusion

We have presented an approach to single sensor source separation based on an extension of Wiener filtering to non-stationary processes, through the use of Gaussian Mixture Models instead of plain Gaussian densities in the standard



(a)

(b)

(c)

(d)

(e)

Fig. 1. from top to bottom : (a) piano+bass source, (b) drum source, (c) mixture of both sources, (d) estimated piano+bass source, (e) estimated drum source

Wiener approach. We have extended the approach to the case of gaussian scaled mixture models, which permits to advantageously separate the PSD shape from the amplitude information.

The presented approach makes use of a preliminary step, in which PSD vectors are estimated on some excerpts of the sources, corresponding to the various GSMM model states. This prior information is needed in order to perform the source separation. Our preliminary experiments show some benefit on the approach as compared to Wiener filtering, on our example.

Many tracks deserve to be further investigated to improve and robustify the proposed approach. For instance, the prior densities that we have used in the Bayesian framework are all phase invariant. Thus we may not recover through these models the true phase of the sources. Phase modelling in the STFT domain should be studied, in order to improve further the approach.

An other step could consist in introducing a psycho-acoustic model (both in the separation step and in the evaluation criteria) in order to optimize the separation in the most perceptible frequency bands for a given source, rather than using a uniform criterion, as is the case in the current approach.

REFERENCES

[1] S. Amari and J. Cardoso. Blind source separation — semiparametric statistical approach. *IEEE Transaction on Signal Processing*, 45(11):2692–2700, December 1997.

[2] A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[3] Elie-Laurent Benaroya. Séparation de plusieurs sources sonores avec un seul microphone. *Thèse de Doctorat de l'Université de Rennes I*, June 2003.

[4] L. Benaroya, R. Gribonval, and F. Bimbot. Non negative sparse representation for wiener based source separation with a single sensor. In *ICASSP*, pages 613–616, Hong Kong, 2003.

[5] Olivier Bermond and Jean-François Cardoso. Approximate likelihood for noisy mixtures. In *Proceedings ICA '99*, pages 325–330, Aussois, France, 1999.

[6] A. Bijaoui. Wavelets, gaussian mixtures and wiener filtering. *Signal Processing*, 82:709–712, 2002.

[7] J.F. Cardoso. Blind signal separation: statistical principles. In *Proceedings of IEEE*, volume 86, pages 2009–2025, 1998.

[8] D. L. Donoho. Denoising by soft-thresholding. *IEEE Transactions on Information Theory*, 41:613–627, May 1995.

[9] Y. Ephraim and N. Merhav. Hidden markov processes. *IEEE Transactions on Information Theory*, 48(6):1518–1569, June 2002.

[10] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte. Proposals for performance measurement in source separation. In *ICA*, pages 715–720, Nara, Japan, 2003.

[11] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13:411–430, 2000.

[12] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. *NIPS*, pages 556–562, 2000.

[13] T. Lee, M. Lewicki, M. Girolami, and T. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, 4(4), April 1999.

[14] 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 *Proceedings of the 8th international conference on Image Processing*, Thessaloniki, Greece, October 2001.

[15] Sam T. Roweis. One microphone source separation. In *NIPS*, pages 793–799, 2000.

[16] M. Tipping and C. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61(3):611–622, 1999.

[17] A.P. Varga and R.K. Moore. Hidden markov model decomposition of speech and noise. In *Proc. ICASSP'90*, pages 845–848, 1990.

[18] S.V. Vaseghi and B.P. Milner. Noise compensation methods for hidden markov model speech recognition in adverse environments. *IEEE Trans. on Speech and Audio Processing*, 5(1):11–21, 1997.

[19] Y. Zhao. Frequency-domain maximum likelihood estimation for automatic speech recognition in additive and convolutive noises. *IEEE Trans. on Speech and Audio Processing*, 8(3):255–266, 2000.