SINGLE CHANNEL SOURCE SEPARATION USING STATIC AND DYNAMIC FEATURES IN THE POWER DOMAIN

Ilyas Potamitis[±] and *Alexey Ozerov*[†]

± Department of Music Technology and Acoustics, Technological Educational Institute of Crete
E. Daskalaki - Perivolia, 74100 Rethymno - Crete, Greece, email: potamitis@stef.teicrete.gr
† TELECOM-ParisTech / CNRS - LTCI, Signal and Image Processing Department
37 rue Dareau, 75014 Paris, France, email: ozerov@telecom-paristech.fr
web: www.wcl.ee.upatras.gr, perso.telecom-paristech.fr/~ozerov/

ABSTRACT

In this paper we address the problem of separating the sound sources composing complex sound mixtures using a single microphone. The a-priori information of static and delta power of each source is represented by Gaussian mixture models (GMMs) and incorporated into a full posterior probability density function. We present a unified probabilistic framework that integrates the a-priori information of the power and the delta power of the sources and we derive a closed-form approximate minimum mean square error (MMSE) estimator of the audio sources. The experimental part evaluates our approach on mixtures of real environmental sounds in scenarios that involve speakers talking in a music background. Comprehensive experiments clarify the importance of incorporating delta in the separation process by presenting separation results using the static only and the joint static and delta a-priori models.

1. INTRODUCTION

Sound events, more often than not, do not appear in isolation but are a part of an audio mixture. Typical real life examples of sums of audio sources are the combination of vocals and instruments in a song, speech in the presence of a noisy background, and overlapping speech in a conversation. The estimation of the contribution of each source to the audio mixture is not a trivial task and it is even harder in the case where only one recording channel of the mixture is available. Most of the state-of-the-art, model-based approaches use the Gaussian mixture models (GMM) or their natural extensions the hidden Markov models (HMM) [1-6] to model probability densities representing the sources. GMM-based approaches allow source separation due to the diversity of source spectral shapes they can represent in their mixtures. In other words, they exploit the fact that the spectra of different sources correspond to the different characteristic spectral patterns of the various combinations that can be observed in the source realizations. The sets of such characteristic spectral patterns constitute essentially the GMM source models.

This work deals with the problem of separating two sources composing audio mixture out of a single recording. The generalization for any number of sources is straightforward. Our approach belongs to the model-based category of single-channel source separation algorithms that model *a-priori* information about the power and the delta power of the sources with GMMs, while the estimation of the independent sources forming the mixture is derived by minimizing an objective function under the Bayesian probabilistic framework. Our formulation follows the line of thought of [3,5,6]. However, there are subtle differences that eventually lead to two novel gain functions. The main differences are:

- First it deduces novel gain functions for the separation of two simultaneous sound sources out of a recording of a single microphone. In this work the mathematical formulation takes place on the power of short-time Fourier transform (STFT) whereas in [3,5,6] the statistical modelling is applied to the complex STFT domain using the concept of complex-valued random vectors.
- In order to deduce an estimation of the sources composing the sound mixture we are using the minimum mean square error (MMSE) of the power spectrum which is closer in spirit to speech/speaker recognition, while in [3,5,6] the MMSE of spectrum is employed.

The dynamic features (deltas and double deltas) have proven themselves as complementary cues of information as they are able to express time correlations between signal frames and are widely used in the automatic speech/speaker recognition (ASR) in addition to the static features enhancing the discriminative power of the recognition process.

This work presents an extension to the problem of single channel source separation by introducing probabilistic models of the time-differences of the power spectrum in the separation thus exploiting the time-dependencies of the signals in the estimation process. By using the dynamic features (in addition to static ones) we demonstrate that the separation performance is improved, as compared to the GMM-based approach applied to the power domain, due to:

- improved discriminative power facilitating correct identification of the characteristic spectral patterns,
- exploiting of time dependencies at the signal level, allowing a better estimation of the spectra of the sources by predicting some parts of them from the previously estimated spectra.

2. PROBABILISTIC SINGLE-CHANNEL SOURCE SEPARATION

2.1 One channel separation in the power domain using static priors only

We present a mathematical formulation for two sources, assuming that a generalization for any number of sources is straightforward. Let $X_{k,t}$ denote a complex domain STFT of the mixture with *k*=1,..,*K* being the frequency-bin index for a fixed-length time window (*K* is the length of the discrete Fourier transform), and *t*=1,..,*T* being the frame index.

Let $S^1_{k,t}$, $S^2_{k,t}$ be the corresponding STFTs of two signal sources. Then

$$X_{k,t} = S^1_{k,t} + S^2_{k,t} \tag{1}$$

Assuming that the sources are independent, the power spectrum of the mixture can be obtained from (1) by ignoring the error in phase:

$$\left|X_{k,t}\right|^2 \simeq \left|S^1_{k,t}\right|^2 + \left|S^2_{k,t}\right|^2 \tag{2}$$

Let $\mathbf{x}_t$, $\mathbf{s}_{1,t}$, $\mathbf{s}_{2,t}$ be the power spectrum vectors, i.e.,

$$\mathbf{x}_t = \begin{bmatrix} \left|X_{1,t}\right|^2 \\ \dots \\ \left|X_{K,t}\right|^2 \end{bmatrix},\ \mathbf{s}_1 = \begin{bmatrix} \left|S^1_{1,t}\right|^2 \\ \dots \\ \left|S^1_{K,t}\right|^2 \end{bmatrix},\ \mathbf{s}_2 = \begin{bmatrix} \left|S^2_{1,t}\right|^2 \\ \dots \\ \left|S^2_{K,t}\right|^2 \end{bmatrix},$$

then (2) becomes

$$\mathbf{x}_t = \mathbf{s}_{1,t} + \mathbf{s}_{2,t} \tag{3}$$

In this work we use the Bayesian statistical framework and we incorporate the *a-priori* information we have for the sources in probability density functions $p(\mathbf{s}_{1,t})$, $p(\mathbf{s}_{2,t})$ represented by Gaussian mixture models (GMM). In other words, we assume that:

$$\begin{aligned} p\left(\mathbf{s}_{1,t}\right) &= \sum_i w_{1,i} N\left(\mathbf{s}_{1,t}; \mu_{1,i}, \Sigma_{1,i}\right) \\ p\left(\mathbf{s}_{2,t}\right) &= \sum_j w_{2,j} N\left(\mathbf{s}_{1,t}; \mu_{2,j}, \Sigma_{2,j}\right) \end{aligned} \tag{4}$$

where $\mu_{1,i}, \mu_{2,j}$ are the mean vectors, $\Sigma_{1,i}, \Sigma_{2,j}$ are the covariance matrices, $w_{1,i} \geq 0, w_{2,j} \geq 0$ are the weights (satisfying $\sum_i w_{1,i} = 1$ and $\sum_j w_{2,j} = 1$), and the subscripts *i, j* are indices running over the mixtures of each source.

A power vector of the mixture is thought to be created by first selecting a mixture number *i* with probability $w_{1,i}$ and then producing an observation following the normal distribution $N\left(\mathbf{s}_{1,t}; \mu_{1,i}, \Sigma_{1,i}\right)$. The same procedure is followed for $\mathbf{s}_{2,t}$ and the results are combined according to (3).

Let $\gamma_{i,j}\left(\mathbf{x}_t\right)$ denote how probable is that the combination *i, j* of states has produced the power of the currently processed mixture frame, i.e.:

$$\gamma_{i,j}\left(\mathbf{x}_t\right) = p\left(i, j \middle| \mathbf{x}_t\right) \propto p\left(\mathbf{x}_t \middle| i, j\right) p\left(i\right) p\left(j\right) \tag{5}$$

If *i, j* are given then from (3) we deduce $p\left(\mathbf{x}_t \middle| i, j\right) = N\left(\mathbf{x}_t; \mu_{1,i} + \mu_{2,j}, \Sigma_{1,i} + \Sigma_{2,j}\right)$, and therefore $\gamma_{i,j}\left(\mathbf{x}_t\right) \propto w_{1,i} w_{2,j} N\left(\mathbf{x}_t; \mu_{1,i} + \mu_{2,j}, \Sigma_{1,i} + \Sigma_{2,j}\right)$.

The optimization criterion for estimating $\mathbf{s}_1$ is MMSE, i.e., $\mathrm{E}\left(\left\|\mathbf{s}_1 - \hat{\mathbf{s}}_1\right\|^2 \middle| \mathbf{x}\right) \xrightarrow[\hat{\mathbf{s}}_1]{} \min$, which is equivalent to take the expectation of power spectrum, thus we estimate the source power spectrum $\mathbf{s}_{1,t}$ (and similarly $\mathbf{s}_{2,t}$) as:

$$\mathrm{E}\left(\mathbf{s}_{1,t} \middle| \mathbf{x}_t\right) = \int \mathbf{s}_{1,t}\, p\left(\mathbf{s}_{1,t} \middle| \mathbf{x}_t\right) d\mathbf{s}_{1,t} \tag{6}$$

$\mathrm{E}\left(\mathbf{s}_{1,t} \middle| \mathbf{x}_t\right)$ can be seen as the result of integrating out of $\mathrm{E}\left(\mathbf{s}_{1,t} \middle| \mathbf{x}_t, i, j\right)$ the hidden states *i, j* and (6) becomes:

$$\mathrm{E}\left(\mathbf{s}_{1,t} \middle| \mathbf{x}_t\right) = \int \mathbf{s}_{1,t} \sum_i \sum_j \gamma_{i,j}(\mathbf{x}_t) p\left(\mathbf{s}_{1,t} \middle| \mathbf{x}_t, i, j\right) d\mathbf{s}_{1,t} \tag{7}$$

which can also be expressed as:

$$\mathrm{E}\left(\mathbf{s}_{1,t} \middle| \mathbf{x}_t\right) = \sum_i \sum_j \gamma_{i,j}(\mathbf{x}_t) \int \mathbf{s}_{1,t}\, p\left(\mathbf{s}_{1,t} \middle| \mathbf{x}_t, i, j\right) d\mathbf{s}_{1,t} \tag{8}$$

By direct application of the Bayes law we have:

$$p\left(\mathbf{s}_{1,t} \middle| \mathbf{x}_t, i, j\right) = \frac{p(\mathbf{x}_t \mid \mathbf{s}_{1,t}, i, j)\, p(\mathbf{s}_{1,t} \mid i, j)}{p(\mathbf{x}_t \mid i, j)} \propto p(\mathbf{x}_t \mid \mathbf{s}_{1,t}, j)\, p(\mathbf{s}_{1,t} \mid i)$$

From (3) we deduce:

$$p(\mathbf{x}_t \mid \mathbf{s}_{1,t}, j)\, p(\mathbf{s}_{1,t} \mid i) = N(\mathbf{x}_t - \mathbf{s}_{1,t}; \mu_{2,j}, \Sigma_{2,j}) N(\mathbf{s}_{1,t}; \mu_{1,i}, \Sigma_{1,i})$$

$$= N(\mathbf{s}_{1,t}; \mathbf{x}_t - \mu_{2,j}, \Sigma_{2,j}) N(\mathbf{s}_{1,t}; \mu_{1,i}, \Sigma_{1,i})$$

By applying lemma 1 from the appendix we conclude that $p\left(\mathbf{s}_{1,t} \middle| \mathbf{x}_t, i, j\right)$ is a Gaussian pdf, i.e.,

$$p\left(\mathbf{s}_{1,t} \middle| \mathbf{x}_t, i, j\right) = N(\mathbf{s}_{1,t}; \hat{\mathbf{s}}^{i,j}_{1,t}, \hat{\Sigma}^{i,j}_{1,t})$$

with

$$\hat{\mathbf{s}}^{i,j}_{1,t} = \left(\Sigma_{1,i} + \Sigma_{2,j}\right)^{-1} \Sigma_{1,i} \left(\mathbf{x}_t - \mu_{2,j}\right) + \left(\Sigma_{1,i} + \Sigma_{2,j}\right)^{-1} \Sigma_{2,j} \mu_{1,i}$$

$$\hat{\Sigma}^{i,j}_{1,t} = \left(\Sigma_{1,i} + \Sigma_{2,j}\right)^{-1} \Sigma_{1,i} \Sigma_{2,j}$$

(see also [7-8] for a situation in robust speech recognition that bears resemblance to the above derivation).

Therefore, we derive a local MMSE estimator $\hat{\mathbf{s}}^{i,j}_{1,t}$ and the covariance $\hat{\Sigma}^{i,j}_{1,t}$ of the estimator as well. The estimator's covariance is useful for estimator's error analysis (see section 3). Finally, from (8) one can directly get the estimator as:

$$\mathrm{E}\left(\mathbf{s}_{1,t} \middle| \mathbf{x}_t\right) = \sum_{i,j} \gamma_{i,j}\left(\mathbf{x}_t\right) \begin{bmatrix} \left(\Sigma_{1,i} + \Sigma_{2,j}\right)^{-1} \Sigma_{1,i} (\mathbf{x}_t - \mu_{2,j}) \\ + \left(\Sigma_{1,i} + \Sigma_{2,j}\right)^{-1} \Sigma_{2,j} \mu_{1,i} \end{bmatrix} \tag{9}$$

2.2 One channel separation in the power domain using static and dynamic priors

We proceed in incorporating the *a-priori* information of dynamic power spectrum in probability density functions for $\mathbf{s}_1,\ \mathbf{s}_2$. The deltas of sources $\mathbf{s}_1,\ \mathbf{s}_2$ are defined as:

$$\Delta\mathbf{s}_{1,t} \equiv \mathbf{s}_{1,t} - \mathbf{s}_{1,t-1},\ \Delta\mathbf{s}_{2,t} \equiv \mathbf{s}_{2,t} - \mathbf{s}_{2,t-1}$$

We define the joint models for $\mathbf{s}_1,\ \Delta\mathbf{s}_1$ and $\mathbf{s}_2,\ \Delta\mathbf{s}_2$ as:

$$p(\mathbf{s}_{1,t}, \Delta\mathbf{s}_{1,t}) = \sum_i w_{1,i} N\left(\mathbf{s}_{1,t}; \mu_{1,i}, \Sigma_{1,i}\right) N\left(\Delta\mathbf{s}_{1,t}; \mu_{\Delta1,i}, \Sigma_{\Delta1,i}\right)$$

$$p\left(\mathbf{s}_{2,t}, \Delta\mathbf{s}_{2,t}\right) = \sum_j w_{2,j} N\left(\mathbf{s}_{2,t}; \mu_{2,j}, \Sigma_{2,j}\right) N\left(\Delta\mathbf{s}_{2,t}; \mu_{\Delta2,j}, \Sigma_{\Delta2,j}\right)$$

where $\mu_{\Delta1,i}, \mu_{\Delta2,j}$ and $\Sigma_{\Delta1,i}, \Sigma_{\Delta2,j}$ are the mean vectors and covariance matrices of deltas.

We proceed on deriving the estimator for $\mathbf{s}_{1,t}$. The procedure is similar for $\mathbf{s}_{2,t}$.

Given estimations of the previous frames $\hat{\mathbf{s}}_{1,t-1},\ \hat{\mathbf{s}}_{2,t-1}$ we estimate the power vector $\mathbf{s}_{1,t}$ by the conditional MMSE as:

$$\hat{\mathbf{s}}_{1,t|t-1} \equiv E\left[\mathbf{s}_{1,t} \middle| \mathbf{x}_t, \hat{\mathbf{s}}_{1,t-1}, \hat{\mathbf{s}}_{2,t-1}\right]$$

Assuming perfect estimators of $\mathbf{s}_{1,t-1},\ \mathbf{s}_{2,t-1}$ we have

$$p\left(\mathbf{s}_{1,t} \middle| \hat{\mathbf{s}}_{1,t-1}, i\right) \simeq p\left(\mathbf{s}_{1,t} \middle| \mathbf{s}_{1,t-1}, i\right), \text{ and}$$

$$p\left(\mathbf{s}_{2,t} \middle| \hat{\mathbf{s}}_{2,t-1}, j\right) \simeq p\left(\mathbf{s}_{2,t} \middle| \mathbf{s}_{2,t-1}, j\right)$$

However,

$$p\left(\mathbf{s}_{1,t} \middle| \mathbf{s}_{1,t-1}, i\right) \propto p\left(\mathbf{s}_{1,t},\ \mathbf{s}_{1,t} - \mathbf{s}_{1,t-1} \middle| i\right) =$$

$$N\left(\mathbf{s}_{1,t}; \mu_{1,i}, \Sigma_{1,i}\right) N\left(\Delta\mathbf{s}_{1,t}; \mu_{\Delta1,i}, \Sigma_{\Delta1,i}\right) =$$

$$N\left(\mathbf{s}_{1,t}; \mu_{1,i}, \Sigma_{1,i}\right) N\left(\mathbf{s}_{1,t}; \mu_{\Delta1,i} + \mathbf{s}_{1,t-1}, \Sigma_{\Delta1,i}\right)$$

By using again lemma 1 we conclude that

$$p\left(\mathbf{s}_{1,t} \middle| \hat{\mathbf{s}}_{1,t-1}, i\right) = N\left(\mathbf{s}_{1,t}; \tilde{\mu}_{1,i}, \tilde{\Sigma}_{1,i}\right), \text{ where}$$

$$\tilde{\mu}_{1,i} = \left(\Sigma_{1,i} + \Sigma_{\Delta1,i}\right)^{-1} \Sigma_{\Delta1,i} \mu_{1,i} + \left(\Sigma_{1,i} + \Sigma_{\Delta1,i}\right)^{-1} \Sigma_{1,i} \left(\hat{\mathbf{s}}_{1,t-1} + \mu_{\Delta1,i}\right)$$

$$\tilde{\Sigma}_{1,i} = \left(\Sigma_{1,i} + \Sigma_{\Delta1,i}\right)^{-1} \Sigma_{1,i} \Sigma_{\Delta1,i} \quad (10)$$

Since $p\left(\mathbf{s}_{1,t} \middle| \hat{\mathbf{s}}_{1,t-1}\right) = \sum_i w_{1,i}\, p\left(\mathbf{s}_{1,t} \middle| \hat{\mathbf{s}}_{1,t-1}, i\right)$ we conclude that

$$p\left(\mathbf{s}_{1,t} \middle| \hat{\mathbf{s}}_{1,t-1}\right) = \sum_i w_{1,i} N\left(\mathbf{s}_{1,t}; \tilde{\mu}_{1,i}, \tilde{\Sigma}_{1,i}\right)$$

$$p\left(\mathbf{s}_{2,t} \middle| \hat{\mathbf{s}}_{2,t-1}\right) = \sum_j w_{2,j} N\left(\mathbf{s}_{2,t}; \tilde{\mu}_{2,j}, \tilde{\Sigma}_{2,j}\right) \quad (11)$$

Thus, we show that conditionally on the previously estimated power spectra the distributions of $\mathbf{s}_{1,t}$ and $\mathbf{s}_{2,t}$ are GMMs with parameters $\{\tilde{\mu}_{1,i}, \tilde{\Sigma}_{1,i}\}_i$ and $\{\tilde{\mu}_{2,j}, \tilde{\Sigma}_{2,j}\}_j$ depending on $\hat{\mathbf{s}}_{1,t-1}$ and $\hat{\mathbf{s}}_{2,t-1}$. Equation (11) is identical to (4) if one replaces $\mu_{1,i}, \mu_{2,j}, \Sigma_{1,i}, \Sigma_{2,j}$ with $\tilde{\mu}_{1,i}, \tilde{\mu}_{2,j}, \tilde{\Sigma}_{1,i}, \tilde{\Sigma}_{2,j}$. Therefore, we can follow directly the inference procedure of section 2.1 and then conclude that the final estimator is:

$$\hat{\mathbf{s}}_{1,t|t-1} = \sum_i \sum_j \gamma_{i,j}(\mathbf{x}_t) \begin{bmatrix} \left(\tilde{\Sigma}_{1,i} + \tilde{\Sigma}_{2,j}\right)^{-1} \tilde{\Sigma}_{1,i} \left(\mathbf{x}_t - \tilde{\mu}_{2,j}\right) \\ + \left(\tilde{\Sigma}_{1,i} + \tilde{\Sigma}_{2,j}\right)^{-1} \tilde{\Sigma}_{2,j} \tilde{\mu}_{1,i} \end{bmatrix} \quad (12)$$

where $\tilde{\mu}_{1,i}$, $\tilde{\Sigma}_{1,i}$ (and similarly for $\tilde{\mu}_{2,j}$, $\tilde{\Sigma}_{2,j}$) are given by (10).

In order to decrease the influence of assuming perfect estimators of $\mathbf{s}_{1,t-1},\ \mathbf{s}_{2,t-1}$ we have applied a penalization factor $r > 1$ of the delta features covariance matrices (see also [7-8] for a similar case). Therefore, (10) becomes:

$$\tilde{\mu}_{1,i} = \left(\Sigma_{1,i} + r\Sigma_{\Delta1,i}\right)^{-1} r\Sigma_{\Delta1,i} \mu_{1,i} + \left(\Sigma_{1,i} + r\Sigma_{\Delta1,i}\right)^{-1} \Sigma_{1,i} \left(\hat{\mathbf{s}}_{1,t-1} + \mu_{\Delta1,i}\right)$$

$$\tilde{\Sigma}_{1,i} = \left(\Sigma_{1,i} + r\Sigma_{\Delta1,i}\right)^{-1} \Sigma_{1,i} r\Sigma_{\Delta1,i}.$$

The experiments presented in Section 4 are for *r*=7, that was found experimentally to give the best results.

3. SINGLE STATE ERROR ANALYSIS

We analyze the separation error in the single state (*i, j*) without considering the influence of the state selection process. Using the expressions for $\tilde{\mu}_{1,i}$ and $\tilde{\Sigma}_{1,i}$ in (10) we can rewrite the expressions for $\hat{\mathbf{s}}^{i,j}_{1,t|t-1}$ and $\hat{\Sigma}^{i,j}_{1,t|t-1}$ (similar to those for $\hat{\mathbf{s}}^{i,j}_{1,t}$ and $\hat{\Sigma}^{i,j}_{1,t}$ from section 2.1) in the following forms:

$$\hat{\Sigma}^{i,j}_{1,t|t-1} = \left(\left(\Sigma_{1,i}\right)^{-1} + \left(\Sigma_{\Delta1,i}\right)^{-1} + \left(\Sigma_{2,j}\right)^{-1} + \left(\Sigma_{\Delta2,j}\right)^{-1}\right)^{-1}$$

$$\hat{\mathbf{s}}^{i,j}_{1,t|t-1} = \hat{\Sigma}^{i,j}_{1,t|t-1} \left(\begin{array}{l} \left(\Sigma_{1,i}\right)^{-1} \mu_{1,i} + \left(\Sigma_{\Delta1,i}\right)^{-1} \left(\hat{\mathbf{s}}_{1,t-1} + \mu_{\Delta1,i}\right) + \\ \left(\Sigma_{2,j}\right)^{-1} \left(\mathbf{x}_t - \mu_{2,j}\right) + \left(\Sigma_{\Delta2,j}\right)^{-1} \left(\mathbf{x}_t - \hat{\mathbf{s}}_{2,t-1} - \mu_{\Delta2,j}\right) \end{array} \right)$$

These expressions have a quite convenient form to analyze the estimator and the covariance of the estimation error.

From the expression for $\hat{\Sigma}^{i,j}_{1,t|t-1}$ we see that in the limit case the smallest covariance dominates in this expression. For example if $\left|\Sigma_{\Delta1,i}\right| << \min\left(\left|\Sigma_{1,i}\right|, \left|\Sigma_{2,j}\right|, \left|\Sigma_{\Delta2,j}\right|\right)$, then $\hat{\Sigma}^{i,j}_{1,t|t-1} \approx \Sigma_{\Delta1,i}$.

Therefore, if we include the dynamic covariances $\Sigma_{\Delta1,i}$ and $\Sigma_{\Delta2,j}$ which are smaller than static covariances $\Sigma_{1,i}$ and $\Sigma_{2,j}$, then $\hat{\Sigma}^{i,j}_{1,t|t-1}$ will be smaller (compared to $\hat{\Sigma}^{i,j}_{1,t|t-1}$ without the dynamic features). This is a case in practice (i.e. $\Sigma_{\Delta1,i}$ is usually smaller than $\Sigma_{1,i}$), since we expect to observe less variation in $\Delta\mathbf{s}_{1,t} = \mathbf{s}_{1,t} - \mathbf{s}_{1,t-1}$ compared to $\mathbf{s}_{1,t}$. Even if $\Sigma_{\Delta1,i}$ and $\Sigma_{\Delta2,j}$ are of the same order as $\Sigma_{1,i}$ and

$\Sigma_{2,j}$, we should still gain by using dynamic features. Indeed, let's assume for example that $\Sigma_{\Delta 1,i} \approx \Sigma_{1,i}$ and $\Sigma_{\Delta 2,j} \approx \Sigma_{2,j}$, then we have:

$\hat{\Sigma}^{i,j}_{1,t|t-1} \approx \frac{1}{2}\left(\left[\Sigma_{1,i}\right]^{-1} + \left[\Sigma_{2,j}\right]^{-1}\right)^{-1}$, which is two times less than the case of static features only (i.e., $\hat{\Sigma}^{i,j}_{1,t|t-1} \approx \left(\left[\Sigma_{1,i}\right]^{-1} + \left[\Sigma_{2,j}\right]^{-1}\right)^{-1}$). The expression for estimate $\hat{\mathbf{s}}^{i,j}_{1,t|t-1}$ is also very easily tractable:

- if $\left|\Sigma_{1,i}\right|$ is small compared to other covariances, then $\hat{\mathbf{s}}^{i,j}_{1,t|t-1} \approx \mu_{1,i}$, i.e., we replace $\mathbf{s}_{1,t}$ by its expected value given by the static model,
- if $\left|\Sigma_{2,j}\right|$ is small compared to other covariances, then $\hat{\mathbf{s}}^{i,j}_{1,t|t-1} \approx \mathbf{x}_t - \mu_{2,j}$, i.e., we estimate $\mathbf{s}_{1,t}$ as a difference between the observed mixture and the expected value of $\mathbf{s}_{2,t}$ given by the static model,
- if $\left|\Sigma_{\Delta 1,i}\right|$ is small compared to other covariances, then $\hat{\mathbf{s}}^{i,j}_{1,t|t-1} \approx \hat{\mathbf{s}}_{1,t-1} + \mu_{\Delta 1,i}$, i.e., we replace $\mathbf{s}_{1,t}$ by its expected value given by the dynamic model (partially predicted using the previous estimate $\hat{\mathbf{s}}_{1,t-1}$),
- if $\left|\Sigma_{\Delta 2,j}\right|$ is small compared to other covariances, then $\hat{\mathbf{s}}^{i,j}_{1,t|t-1} \approx \mathbf{x}_t - \hat{\mathbf{s}}_{2,t-1} - \mu_{\Delta 2,j}$.

Therefore we conclude that the joint employment of static and dynamic components should perform better than using static *a-priori* information only. With GMM-based approach we always estimate our sources based on some selected codeword (characteristic spectral shape) of a fixed codebook (GMM). With the incorporation of GMM that model the deltas we alternate between prediction from the selected codeword and prediction from the previously estimated spectrum. The choice between these two options is done so that the estimation variance is minimized. Thus, with the use of delta spectrum we have a smaller covariance of the estimation error, as compared to the GMM-based approach; therefore, we expect achieving a better separation performance.

4. SEPARATION EXPERIMENTS

The experimental setup is focused on separating speech from background music. The music dataset consists of 10 pieces of instrumental music (each piece is about 2 min long). The speech dataset consists of two sets of recordings: 100 speech sentences from TIMIT database (10 distinct male speakers) and 100 sentences of female speakers, so that we have 10 sentences per speaker.

We trained a male and a female speech model. The 100 files of each set are partitioned in 4 subsets of 25 files each forming a validation set. The training set corresponding to each validation subset is composed of the recordings of the rest of the subsets (4 subsets of 75 recordings). For music model training we use the 60 last seconds of the corresponding music piece. Each speech recording of the evaluation database is mixed with each file of the 10 music recordings, therefore the evaluation datatabase is made by a total of 4*25*10=1000 mixture recordings of male speech and music and 1000 mixtures of female speech and music.

Each recording is downsampled to 11025 Hz. The STFT is computed using half-overlapped 512 samples (46 ms) length Hann windows, and the inverse STFT is computed using rectangular windows. For the power domain separation methods (where only the magnitudes of source spectra are estimated) the phases of the mixture are used for re-synthesis of estimated sources in time domain [3,5,6]. Each GMM model is composed of 16 states with diagonal covariance matrices. The models are initialized using 5 iterations of the K-means algorithm and trained using a standard version of the expectation-maximization algorithm [9] (see e.g. [5] for more details). The measures that are used to assess the proposed static and static+dynamic versions are two commonly used for source separation [10]:

1. Signal to distortion ratio (SDR) defined as:

$$10\log_{10}\frac{\left\langle \hat{s}_m, s_m \right\rangle^2}{\left\|\hat{s}_m\right\|^2 \left\|s_m\right\|^2 - \left\langle \hat{s}_m, s_m \right\rangle^2}$$

   where, s_m is the true reference source signal and $\hat{s}_m$ is the estimated one and *m*=1,2.
2. Segmental SDR (SegSDR): we compute SDR for short segments of 512 samples overlapping by 50%, convert them in dB and average it over the total sequence. SegSDR is sensitive to local separation errors, thus allowing discovering them.

The experimental part aims to explore the benefit of incorporating the dynamic features into the separation process. The proposed versions of static and static+dynamic power features are also compared against a state-of-the-art algorithm [3,5] achieving at least comparable separation results. However, the method from [3,5] slightly outperforms the proposed methods. We think that this is due to the fact that the spectral MMSE [3,5] is more consistent with separation performance measures employed (SDR and SegSDR) than the proposed power spectrum MMSE [6]. The incorporation of the dynamic features results into a small but consistent gain in the separation performance outperforming the static-only case as predicted by the theoretical analysis and as demonstrated by the objective measures depicted in Tables 1a-1b. The results are the mean scores over 1000 mixture recordings for the male case and 1000 files for the female case. However we are currently looking into different versions of dynamic features in order to enhance this gain. Three separation techniques are compared: the static power case (stat), the static and dynamic features (stat+dynam), and a state of the art one-channel separation method (spectral [3,5]). In bold letters are the best results of all three methods compared. In italics, are the best results of the comparison of the power static and static+dynamic cases.

	Male speech		Female speech	
	SDR	**SegSDR**	**SDR**	**SegSDR**
Spectral	9,25	**2,83**	8,89	**2,74**
Stat	9,12	2,31	9,06	2,31
Stat+Dynam	***9,34***	*2,53*	***9,38***	*2,58*

Table 1a) Mean separation results for speech+music mixtures. The evaluation set of male and female speech signals as extracted from the mixture is compared to the corresponding original recordings used to make the mixture.

	Music male mix		Music female mix	
	SDR	**SegSDR**	**SDR**	**SegSDR**
Spectral	**6,16**	**7,55**	**5,64**	**7,48**
Stat	5,33	7,06	4,76	6,89
Stat+Dynam	*5,61*	*7,44*	*5,18*	*7,18*

Table 1b) Mean separation results for speech+music mixtures. The evaluation set of the music part as extracted from the mixture is compared to the corresponding original music recordings.

5. CONCLUSIONS

The aim of this work is to introduce a unifying probabilistic framework for one-channel separation using static only and joint static and dynamic power information. We have derived two novel closed-form estimators of signal separation in the power domain. The use of deltas as in the case of automatic speech/speaker recognition has proven to supply complementary information to the static features as regards the separation task. The algorithms can, in principle, separate an arbitrary number of audio sources out of a single mixture provided that there are available probabilistic descriptions of each source in the form of GMMs.

Online separation samples of the spectral, static and static plus dynamic techniques are provided for subjective evaluation by the readers at:
http://perso.telecom-paristech.fr/~ozerov/demos.html

ACKNOWLEDGMENT

This work was supported in part by the E.C under the 7th framework grant Prometheus 214901.

APPENDIX

Lemma 1 *If the density of a random vector* $\mathbf{s}$ *is proportional to the product of two Gaussian densities, i.e.,*

$$p(\mathbf{s}) = c \cdot N(\mathbf{s}; \mu_1, \Sigma_1) N(\mathbf{s}; \mu_2, \Sigma_2)$$

(c *is a normalization constant), then* $\mathbf{s}$ *is a Gaussian random vector as well (i.e.,* $p(\mathbf{s}) = N(\mathbf{s}; \hat{\mu}, \hat{\Sigma})$*) with mean vector* $\hat{\mu}$ *and covariance matrix* $\hat{\Sigma}$ *defined as :*

$$\hat{\mu} = \left(\Sigma_1 + \Sigma_2\right)^{-1} \Sigma_1 \mu_2 + \left(\Sigma_1 + \Sigma_2\right)^{-1} \Sigma_2 \mu_1$$

$$\hat{\Sigma} = \left(\Sigma_1 + \Sigma_2\right)^{-1} \Sigma_1 \Sigma_2 .$$

Proof If one completes the squares of the Gaussians and rearranges the terms of $p(\mathbf{s})$ one can see that $p(\mathbf{s}) = N(\mathbf{s}; \hat{\mu}, \hat{\Sigma})$. In order to find $\hat{\mu}$ and $\hat{\Sigma}$, we subsequently introduce the following auxiliary function:

$$\Phi(\mathbf{s}) = \log N(\mathbf{s}; \hat{\mu}, \hat{\Sigma}) =$$

$$\log N(\mathbf{s}; \mu_1, \Sigma_1) + \log N(\mathbf{s}; \mu_2, \Sigma_2) + \log c$$

From the definition of Gaussian distribution one can easily figure out that $\hat{\mu}$ is the solution of $\partial / \partial \mathbf{s}\, \Phi(\mathbf{s}) = 0$ and that $\hat{\Sigma} = -\left[\frac{\partial^2}{\partial \mathbf{s} \partial \mathbf{s}} \Phi(\mathbf{s})\right]^{-1}$. By computing the first and the second derivatives of $\Phi(\mathbf{s})$ we obtain the above-mentioned expressions for $\hat{\mu}$ and $\hat{\Sigma}$.

REFERENCES

[1] S. Roweis, "One microphone source separation", in Proc. NIPS, pp. 793-799, 2000.

[2] T. Kristjansson, H. Attias, and J. Hershey, "Single microphone source separation using high resolution signal reconstruction", in Proc. ICASSP, pp. 817-820, 2004.

[3] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor", IEEE Transactions on Speech and Audio Processing, vol.14, no.1, pp. 191-199, Jan. 2006.

[4] D. Ellis, "Model-Based Scene Analysis", Chap. 4 of Computational Auditory Scene Analysis: Principles, Algorithms, and Applications, D. Wang & G. Brown, eds., Wiley/IEEE Press, pp. 115-146, 2006.

[5] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs", IEEE Trans on Audio, Speech, and Language processing, pp. 1564-1578, vol. 15, no. 5, 2007.

[6] A. Ozerov, R. Gribonval, P. Philippe, and F. Bimbot, "Choix et adaptation des modèles pour la séparation de voix chantée à partir d'un seul microphone", Traitement du signal, vol. 24, no. 3, pp. 211-224, 2007.

[7] L. Deng, J. Droppo, and A. Acero, "Estimating Cepstrum of Speech Under the Presence of Noise Using a Joint Prior of Static and Dynamic Features", IEEE Transactions on Speech and Audio Processing, pp. 218-233, vol. 12, no. 3, 2004.

[8] 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 Transactions on Speech and Audio Processing, vol.13, no.3, pp. 412-421, May 2005.

[9] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, Series B, 39(1), pp. 1–38, 1977.

[10] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in Blind Audio Source Separation", IEEE Trans. Audio, Speech and Language Processing, vol. 14, no. 4, pp. 1462-1469, Jul. 2006.