ONE MICROPHONE SINGING VOICE SEPARATION USING SOURCE-ADAPTED MODELS

Alexey Ozerov, Pierrick Philippe

France Télécom R&D
4, rue du Clos Courtel, BP 91226,
35512 Cesson Sévigné cedex, France
alexey.ozerov@francetelecom.com,
pierrick.philippe@francetelecom.com

Rémi Gribonval, Frédéric Bimbot

IRISA (CNRS & INRIA) - projet METISS
Campus de Beaulieu,
35042 Rennes Cedex, France
remi.gribonval@irisa.fr,
frederic.bimbot@irisa.fr

ABSTRACT

In this paper, the problem of one microphone source separation applied to singing voice extraction is studied. A probabilistic approach based on Gaussian Mixture Models (GMM) of the short time spectra of two sources is used. The question of source model adaptation is investigated in order to improve separation quality. A new adaptation method consisting in a filter adaptation technique via the Maximum Likelihood Linear Regression (MLLR) is presented with an associated filter-adapted training phase.

1. INTRODUCTION

The problem of one microphone source separation [1] is a challenging task. In this paper, this problem is studied in the case of singing voice extraction from mono audio recordings. The approach is based on *a priori* probabilistic models for two sources: one being the voice to be extracted from the recording, the second source being the background music. It is assumed that each recording $x(n)$ (called *mixture*) is a simple sum of a voice signal $v(n)$ and a music signal $m(n)$ (called *sources*), where n is a discrete time index ($x(n) = v(n) + m(n)$). The aim is to estimate the voice contribution $\hat{v}(n)$ in the observed signal $x(n)$.

For speech enhancement [2] and separation of several sources in a monophonic musical recording [3] it has been proposed to model the short time spectra of the sources by Gaussian Mixture Models (GMM). These models are learned from training sources. The performance obtained with *general models*, i.e., models learned on training sources issued from recordings different from those to be separated, is rather poor. In the case of our task, large sound classes (voice and music) should be modeled. It may be more efficient to use *adapted models*, i.e., models with characteristics close to those of the mixed sources.

For blind clustering of popular music, Tsai [4] proposes to adapt music and voice models directly from the recording. In a first phase each recording is automatically segmented in a succession of vocal and non-vocal parts. Then, an adapted music model is learned on the non-vocal parts. Finally, using the adapted music model as an *a priori*, an adapted voice model is learned from the vocal parts. Notice that the singing voice does not appear alone, but polluted with background music. Thus, for correct voice model adaptation this background music is attenuated using the adapted music model.

The first part of our contribution consists in the application of this adaptation technique for the singing voice extraction task with some modifications. Secondly, a new solution for voice model adaptation is proposed. This new solution assumes that the adapted voice model is obtained from the general voice model by a linear transformation of the feature space (short time spectra). In that case the transformation is a linear filter, which estimation (referred as *filter adaptation*) is based on the MLLR framework [5]. A filter-adapted training procedure for a general voice model is also presented.

The paper is organized in the following way. The GMM-based one microphone source separation technique [2, 3] is described in section 2. In section 3, a technique of model adaptation is presented, which is based on a segmentation of the mixture into vocal and non-vocal parts. In section 4 the filter adaptation method and filter-adapted training procedure are introduced. The experimental conditions and simulation results are given in section 5.

2. GMM-BASED SOURCE SEPARATION

In this section we recall the principles of GMM-based source separation [2, 3]. The separation scheme is represented in figure 1. We first recall the notion of GMM and explain how they are used to perform adapted Wiener filtering. Eventually, we explain how GMM are learned from training data.

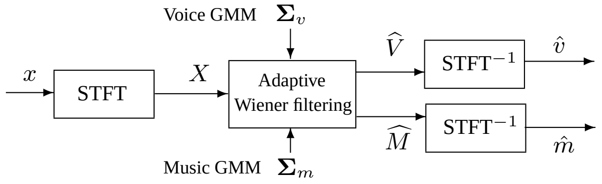
Voice GMM $\mathbf{\Sigma}_v$
x
STFT
X
Adaptive Wiener filtering
$\widehat{V}$
$\widehat{M}$
STFT^{-1}
STFT^{-1}
$\hat{v}$
$\hat{m}$
Music GMM $\mathbf{\Sigma}_m$

Figure 1: GMM-based separation scheme.

2.1. GMM sources modeling

The short time Fourier spectra V_t at time t of the voice signal v are modeled with a GMM, i.e., the probability density function of V_t is given by

$$p(V_t|\mathbf{\Sigma}_v) = \sum_i \omega_{vi} N(V_t; \Sigma_{vi}), \tag{1}$$

with $N(V_t;\Sigma_{vi}) = \prod_f \left[\pi^{-1}\sigma_{vi}^{-2}(f)\exp\left(-|V_t(f)|^2/\sigma_{vi}^2(f)\right)\right]$, where $V_t(f)$ is the complex value of the short time Fourier spectrum V_t at frequency f and $\sigma_{vi}^2(f)$, representing the local Power Spectral Density (PSD) at frequency f in the state i of the GMM, is the diagonal element of the diagonal covariance matrix Σ_{vi}. This GMM is denoted $\boldsymbol{\Sigma}_v = \{\omega_{vi}, \Sigma_{vi}\}_i$. Similarly M_t is modeled by a GMM $\boldsymbol{\Sigma}_m = \{\omega_{mj}, \Sigma_{mj}\}_j$.

2.2. Separation by adaptive Wiener filtering

Source separation is performed in the Short Time Fourier Transform (STFT) domain with the Minimum Mean Square Error (MMSE) estimator, which can be viewed as a form of adaptive Wiener filtering:

$$\begin{aligned}\widehat{V}_t(f) &\triangleq E[V_t(f)|X_t, \boldsymbol{\Sigma}_v, \boldsymbol{\Sigma}_m] \\ &= \sum_{i,j} \gamma_{ij}(t) \frac{\sigma_{vi}^2(f)}{\sigma_{vi}^2(f)+\sigma_{mj}^2(f)} X_t(f),\end{aligned} \qquad (2)$$

with $\sum_{i,j}\gamma_{ij}(t) = 1$ and

$$\begin{aligned}\gamma_{ij}(t) &\triangleq P(q_{vt}=i, q_{mt}=j|X_t, \boldsymbol{\Sigma}_v, \boldsymbol{\Sigma}_m) \\ &\propto \omega_{vi}\omega_{mj} N(X_t; \Sigma_{vi}+\Sigma_{mj}),\end{aligned} \qquad (3)$$

where X_t is the short time Fourier spectrum of the mixture x and q_{vt} and q_{mt} are hidden states of the models $\boldsymbol{\Sigma}_v$ and $\boldsymbol{\Sigma}_m$ at the time t. The time domain source estimation $\hat{v}_n$ is calculated as the inverse STFT of $\widehat{V} = \{\widehat{V}_t\}_t$.

2.3. Model learning

The models $\boldsymbol{\Sigma}_v$ and $\boldsymbol{\Sigma}_m$ are learned by maximization of the likelihoods $p(\bar{V}|\boldsymbol{\Sigma}_v)$ and $p(\bar{M}|\boldsymbol{\Sigma}_m)$, given $\bar{V}$ and $\bar{M}$ the STFT of the training signals. This maximization is achieved using the Expectation Maximization (EM) algorithm [6] initialized by Vector Quantization (VQ). For example, in the case of voice model estimation, the *observed data* $\eta = \bar{V}$ is completed by the *latent data* $\theta = q_v$ (states sequence), and the model parameters $\xi = \boldsymbol{\Sigma}_v$ are estimated by EM which is an iterative algorithm based on the two following steps:

$$\begin{aligned}&\text{Expectation:} && Q(\xi, \xi^{(l)}) = E_\theta\left[\log p(\eta, \theta|\xi)|\eta, \xi^{(l)}\right] \\ &\text{Maximization:} && \xi^{(l+1)} = \arg\max_\xi Q(\xi, \xi^{(l)})\end{aligned} \qquad (4)$$

where $\xi^{(l)}$ denotes the model parameters estimated at the l-th iteration.

3. MODEL ADAPTATION

Let $\mathbf{voc}$ denote the indices of the frames where voice is present in X. Motivated by [4] we learn the music model $\boldsymbol{\Sigma}_m$ from the non-vocal frames $(X_t)_{t\notin\mathbf{voc}} = (M_t)_{t\notin\mathbf{voc}}$ and then estimate the voice model $\boldsymbol{\Sigma}_v$ from the vocal frames in a maximum likelihood manner as follows:

$$\boldsymbol{\Sigma}_v^* = \arg\max_{\boldsymbol{\Sigma}_v} p((X_t)_{t\in\mathbf{voc}}|\boldsymbol{\Sigma}_v, \boldsymbol{\Sigma}_m) \qquad (5)$$

The adaptation procedure is represented in figure 2. In practice, the problem (5) is also solved by EM with observed data $\eta = (X_t)_{t\in\mathbf{voc}}$, latent data $\theta = \{q_v, q_m, (V_t)_{t\in\mathbf{voc}}\}$ and model parameters $\xi = \boldsymbol{\Sigma}_v$, leading in the case of our GMM models to the following re-estimation equations [7]:

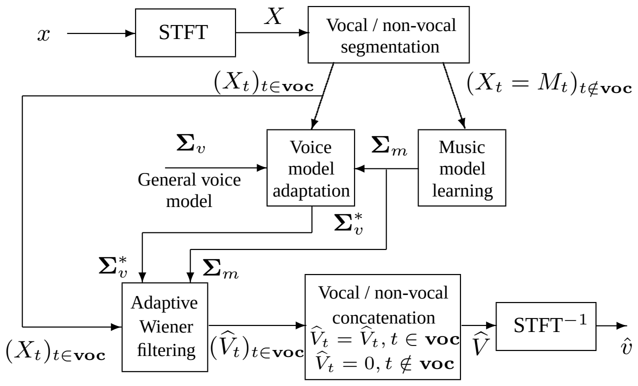
x
STFT
X
Vocal / non-vocal segmentation
$(X_t)_{t\in\mathbf{voc}}$
$(X_t = M_t)_{t\notin\mathbf{voc}}$
$\boldsymbol{\Sigma}_v$
General voice model
Voice model adaptation
$\boldsymbol{\Sigma}_m$
Music model learning
$\boldsymbol{\Sigma}_v^*$
$\boldsymbol{\Sigma}_v^*$
$\boldsymbol{\Sigma}_m$
Adaptive Wiener filtering
$(X_t)_{t\in\mathbf{voc}}$
$(\widehat{V}_t)_{t\in\mathbf{voc}}$
Vocal / non-vocal concatenation
$\widehat{V}_t = \widehat{V}_t, t\in\mathbf{voc}$
$\widehat{V}_t = 0, t\notin\mathbf{voc}$
$\widehat{V}$
STFT^{-1}
$\hat{v}$

Figure 2: Source-adapted separation scheme.

$$\omega_{vi}^{(l+1)} = \frac{1}{T_{\mathbf{voc}}} \sum_{t\in\mathbf{voc}} \sum_j \gamma_{ij}^{(l)}(t), \qquad (6)$$

$$[\sigma_{vi}^{(l+1)}(f)]^2 = \frac{\sum_{t\in\mathbf{voc}}\sum_j \gamma_{ij}^{(l)}(t)\left\langle |V_t(f)|^2\right\rangle_{ij}^{(l)}}{\sum_{t\in\mathbf{voc}}\sum_j \gamma_{ij}^{(l)}(t)}, \qquad (7)$$

$$\begin{aligned}&\left\langle |V_t(f)|^2\right\rangle_{ij}^{(l)} \triangleq E\left[|V_t(f)|^2 \,\middle|\, X_t, q_{vt}=i, q_{mt}=j, \boldsymbol{\Sigma}_v^{(l)}, \boldsymbol{\Sigma}_m\right] \\ &= \frac{[\sigma_{vi}^{(l)}(f)]^2\sigma_{mj}^2(f)}{[\sigma_{vi}^{(l)}(f)]^2+\sigma_{mj}^2(f)} + \left|\frac{[\sigma_{vi}^{(l)}(f)]^2}{[\sigma_{vi}^{(l)}(f)]^2+\sigma_{mj}^2(f)} X_t(f)\right|^2,\end{aligned} \qquad (8)$$

where $T_{\mathbf{voc}}$ is the number of the vocal frames and $\gamma_{ij}^{(l)}(t)$ are computed as in (3). The adaptation algorithm is initialized using a learned general voice model, i.e., $\boldsymbol{\Sigma}_v^{(0)} = \boldsymbol{\Sigma}_v$.

4. FILTER-INVARIANT MODELING

There are a lot of variability factors between the singing voices in different recordings of the collection from which the general voice model is learned in the previous approach. In particular, since each recording might be captured with a specific microphone, in a room with its specific acoustics, there are sources of variability between recordings that can be modeled by a global causal linear time-invariant filter. Instead of building GMM where many Gaussian states are spent modeling the inter-recording variability, we propose to use the states more to model the internal dynamics of the "generic" vocal source, introducing a filter-invariant modeling.

4.1. Voice modeling

With this purpose, we model each voice recording v_r as a convolution $v_r = h_r \star o_r$ with h_r a global filter and o_r an "*original voice*". The short time spectra of the original voices o_r are now modeled by the same GMM $\mathbf{\Sigma}_v$ shared between different recordings, but each recording has its own filter h_r. In the STFT domain this convolution becomes approximately:

$$V_{rt}(f) = |H_r(f)|\tilde{O}_{rt}(f), \quad \tilde{O}_{rt}(f) \triangleq O_{rt}(f)\frac{H_r(f)}{|H_r(f)|}, \tag{9}$$

where O_r and H_r stand for the STFT of o_r and h_r. The short time spectra $\tilde{O}_{rt}$ are modeled by the GMM $\mathbf{\Sigma}_v$ and it is clear from (9) that the probability density of the recorded voice STFT V_r is that of the GMM model

$$\mathbf{\Sigma}_{vr} = \mathcal{H}_r \mathbf{\Sigma}_v \triangleq \{\omega_{vi}, \mathcal{H}_r \Sigma_{vi}\}_i \tag{10}$$

with $\mathcal{H}_r \triangleq \mathrm{diag}[|H_r(f)|^2]_f$.

4.2. Filter adaptation via MLLR

To use this new model at the separation stage on a new recording X, the full adaptation of all voice model parameters (5) is replaced by the only adaptation of the global filter (see figure 3):

$$\mathcal{H}^* = \arg\max_{\mathcal{H}} p((X_t)_{t\in\mathbf{voc}}|\mathcal{H}\mathbf{\Sigma}_v, \mathbf{\Sigma}_m) \tag{11}$$

Such a maximum likelihood estimation corresponds to the Maximum Likelihood Linear Regression (MLLR) framework [5].

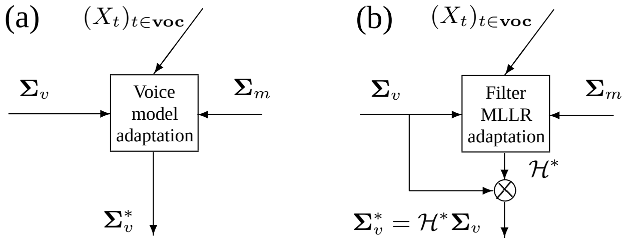

Figure 3: Two voice model adaptation approaches. (a): full model adaptation (b): filter adaptation via MLLR.

Applying of the EM algorithm (4) with observed data $\eta = (X_t)_{t\in\mathbf{voc}}$, latent data $\theta = \{q_v, q_m, (V_t)_{t\in\mathbf{voc}}\}$ and model parameters $\xi = \mathcal{H}$ yields the re-estimation equation:

$$|H^{(l+1)}(f)|^2 = \frac{1}{T_{\mathbf{voc}}} \sum_{t\in\mathbf{voc}} \sum_i \frac{\sum_j \left\langle |V_t(f)|^2 \right\rangle_{ij}^{(l)} \gamma_{ij}^{(l)}(t)}{\sigma_{vi}^2(f)}, \tag{12}$$

where $\left\langle |V_t(f)|^2 \right\rangle_{ij}^{(l)}$ and $\gamma_{ij}^{(l)}(t)$ are calculated as in (8) and (3), replacing the model $\mathbf{\Sigma}_v$ by $\mathcal{H}^{(l)}\mathbf{\Sigma}_v$. The mathematical development of equation (12) is similar to [7].

4.3. Filter-adapted training

The above filter adaptation technique is also applied to general voice model training. The general voice model $\mathbf{\Sigma}_v$ and the unknown filters $\mathbf{H} = \{\mathcal{H}_r\}_r$ are jointly estimated as follows:

$$(\mathbf{H}^*, \mathbf{\Sigma}_v^*) = \arg\max_{(\mathbf{H},\mathbf{\Sigma}_v)} \prod_r p(\bar{V}_r|\mathcal{H}_r\mathbf{\Sigma}_v) \tag{13}$$

where $\bar{\mathbf{V}} = \{\bar{V}_r\}_r$ denotes the STFT of the training recordings with singing voice.

It is difficult to directly apply the EM algorithm (4) with observed data $\eta = \bar{\mathbf{V}}$, latent data $\theta = q_v$ and estimated parameters $\xi = \{\mathbf{H}, \mathbf{\Sigma}_v\}$ to solve the problem (13), since the maximization step is not easy to solve jointly on $\mathbf{\Sigma}_v$ and $\mathbf{H}$. Instead a version of Space-Alternating Generalized EM (SAGE) algorithm [8] is used. The set of estimated parameters ξ is split in two parts $\xi_1 = \mathbf{H}$ and $\xi_2 = \mathbf{\Sigma}_v$. The iteration number $l+1$ of this algorithm consists in two EM algorithm iterations (4). The first iteration is applied to update ξ_1 with $\xi_2 = \xi_2^{(l)}$ fixed and the second one to update ξ_2 with $\xi_1 = \xi_1^{(l+1)}$ fixed. This leads to the following re-estimation equations:

- First EM iteration ($\mathbf{H}$ updated, $\mathbf{\Sigma}_v = \mathbf{\Sigma}_v^{(l)}$ fixed):

$$|H_r^{(l+1)}(f)|^2 = \frac{1}{T_r} \sum_{t=1}^{T_r} \sum_i \frac{|\bar{V}_{rt}(f)|^2}{[\sigma_{vi}^{(l)}(f)]^2} \gamma_{ri}^{(l)}(t), \tag{14}$$

where $\gamma_{ri}^{(l)}(t) \propto \omega_{vi}^{(l)} N(\bar{V}_{rt}; \mathcal{H}_r^{(l)}\Sigma_{vi}^{(l)})$, $\sum_i \gamma_{ri}^{(l)}(t) = 1$ and T_r denotes the number of frames in the STFT for the r-th recording.

- Second EM iteration ($\mathbf{\Sigma}_v$ updated, $\mathbf{H} = \mathbf{H}^{(l+1)}$ fixed):

$$\omega_{vi}^{(l+1)} = \frac{1}{\sum_r T_r} \sum_r \sum_{t=1}^{T_r} \tilde{\gamma}_{ri}^{(l)}(t), \tag{15}$$

$$[\sigma_{vi}^{(l+1)}(f)]^2 = \frac{\sum_r \sum_{t=1}^{T_r} \tilde{\gamma}_{ri}^{(l)}(t) \frac{|\bar{V}_{rt}(f)|^2}{|H_r^{(l+1)}(f)|^2}}{\sum_r \sum_{t=1}^{T_r} \tilde{\gamma}_{ri}^{(l)}(t)}, \tag{16}$$

where $\tilde{\gamma}_{ri}^{(l)}(t) \propto \omega_{vi}^{(l)} N(\bar{V}_{rt}; \mathcal{H}_r^{(l+1)}\Sigma_{vi}^{(l)})$, $\sum_i \tilde{\gamma}_{ri}^{(l)}(t) = 1$.

It should be noted, that during the adapted music model learning, which is performed on the non-vocal parts, the corresponding filter is implicitly adapted. As a matter of fact, there is no need to explicitly make the filter adaptation.

5. RESULTS

This section includes the experimental data description, presentation of the performance measure and the simulation results.

5.1. Data description

The training data base for the general voice model includes 34 samples of singing men's voices from popular music. Each sample is approximately one minute long. The general music model is trained on 30 samples of popular music free from voice. Each sample is also about one minute long and all samples come from different artists. The test database contains five songs of the same genre, for which the voice and music tracks are available separately. It is therefore possible to evaluate the separation performance by comparing the estimated voice with the original one. The test items are manually segmented in vocal and non-vocal parts.

Since state of the art single channel separation techniques (including ours) so far only provide rather low quality sounds, we have chosen to work with recordings made at a rather low sampling frequency of 11025 Hz. This seemed to be a good trade-off between quality and computational complexity.

5.2. Performance measure

To measure the quality of the estimation $\hat{v}$ with respect to the original singing voice v, we use the Source to Distortion Ratio (SDR) calculated as follows [9]:

$$\mathrm{SDR}(\hat{v}, v) = 10 \log_{10} \left[\frac{\langle \hat{v}, v \rangle^2}{\|\hat{v}\|^2 \|v\|^2 - \langle \hat{v}, v \rangle^2} \right] \qquad (17)$$

where $\langle \hat{v}, v \rangle$ is the scalar product of $\hat{v}$ and v, $\|v\|^2$ is the energy of v. To evaluate the separation performance for one recording, the Normalized SDR (NSDR) is used, it measures the improvement of the SDR between the non-processed mixture x and the estimated voice $\hat{v}$: $\mathrm{NSDR}(\hat{v}, x, v) = \mathrm{SDR}(\hat{v}, v) - \mathrm{SDR}(x, v)$. For overall performance estimation the Global NSDR (GNSDR) is calculated averaging the NSDR over different recordings.

5.3. Simulations

The simulations are performed for different combinations of 32-states voice GMM and 32-states music GMM in order to show the effect of different adaptation steps. The STFT is calculated using the half-overlapped 93-ms length Hamming windows. The separation is only made on the vocal parts. The simulation results are represented in table 1.

Voice model	Music model	GNSDR (dB)
$\mathbf{\Sigma}_v^G[\bar{\mathbf{V}}]$	$\mathbf{\Sigma}_m^G[\bar{\mathbf{M}}]$	5.06
$\mathbf{\Sigma}_v^G[\bar{\mathbf{V}}]$	$\mathbf{\Sigma}_m^A[(X_t)_{t \notin \mathbf{voc}}]$	9.09
$\mathcal{H}^G[(X_t)_{t \in \mathbf{voc}}]\mathbf{\Sigma}_v^G[\bar{\mathbf{V}}]$	$\mathbf{\Sigma}_m^A[(X_t)_{t \notin \mathbf{voc}}]$	9.81
$\mathcal{H}^F[(X_t)_{t \in \mathbf{voc}}]\mathbf{\Sigma}_v^F[\bar{\mathbf{V}}]$	$\mathbf{\Sigma}_m^A[(X_t)_{t \notin \mathbf{voc}}]$	10.05
$\mathbf{\Sigma}_v^{\mathrm{Ref}}[(V_t)_{t \in \mathbf{voc}}]$	$\mathbf{\Sigma}_m^A[(X_t)_{t \notin \mathbf{voc}}]$	*12.54*

Table 1: Simulation results. The data used for model / filter training is given in the braces.

The first experiments used a general voice model $\mathbf{\Sigma}_v^G$ and a general music model $\mathbf{\Sigma}_m^G$ learned from the voice training data $\bar{\mathbf{V}}$ and music training data $\bar{\mathbf{M}}$.

Learning the adapted music model $\mathbf{\Sigma}_m^A$ from the non-vocal parts of each testing song $(X_t)_{t \notin \mathbf{voc}}$ increases the GNSDR by about 4 dB in comparison with the general music model $\mathbf{\Sigma}_m^G$.

The overall performance is again increased by about 0.7 dB when a filter is adapted on the vocal parts for the same voice model $\mathbf{\Sigma}_v^G$ (see eq. (11)).

A slight gain about 0.25 dB is observed when the voice model $\mathbf{\Sigma}_v^F$ used in the filter-adapted separation is learned using the filter-adapted training procedure (see sec. 4.3).

For comparison we also computed an empirical performance upper bound using a reference voice model $\mathbf{\Sigma}_v^{\mathrm{Ref}}$ learned from the vocal parts of the singing voice track alone. These tracks, which are not accessible in a real setting, are here available for evaluation purposes.

Our proposal is based on:

- music model learning on the non-vocal parts
- filter-adapted learning of the general voice model
- filter adaptation of the voice model at the separation stage

Compared to the use of non-adapted models, it brings a fair 5 dB improvement and it remains only 2.5 dB below the empirical performance bound.

The full voice model adaptation technique described in section 3 has also been tested. The adapted voice model is obtained with the EM algorithm (6 - 8) initialized by the general voice model $\mathbf{\Sigma}_v^F$ (the initialization by VQ from $(X_t)_{t \in \mathbf{voc}}$ has also been tested giving a quite bad result: GNSDR = 2.64 dB). The GNSDR = 9.9 dB is obtained, which is quite close to the result with the MLLR filter adaptation technique (10.05 dB). However, it has been noticed that in contrast to the MLLR adaptation procedure, the full voice model adaptation technique sometimes leads to certain listening impairments.

6. CONCLUSION

In the context of one microphone source separation applied to singing voice extraction, the question of adaptation of the *a priori* source models has been studied, in the case where a song is already segmented into vocal and non-vocal parts. The new MLLR filter adaptation technique for voice model adaptation is proposed together with a filter-adapted training procedure. The simulation results show that each adaptation step leads to improvement of the separation performance. The MLLR filter adaptation method is compared with the full voice model adaptation technique.

It should be noted that in comparison with the state of the art approaches [1, 3], where the training sources similar to those to be separated are needed to achieve a satisfactory separation performance, our framework can still be applied in real conditions, since the manual vocal / non-vocal segmentation of songs can be made by the user. In the future we are going to replace this manual segmentation by an automatic segmentation module.

7. REFERENCES

[1] Sam T. Roweis. One microphone source separation. In *Advances in Neural Information Processing Systems 13*, pages 793–799. MIT Press, 2001.

[2] Y. Ephraim. A bayesian estimation approach for speech enhancement using hidden markov models. *IEEE Trans. Signal Processing*, SP-40:725–735, April 1992.

[3] L. Benaroya and F. Bimbot. Wiener based source separation with HMM/GMM using a single sensor. In *ICA*, 2003.

[4] Wei-Ho Tsai, Dwight Rogers and Hsin-Min Wang. Blind clustering of popular music recordings based on singer voice characteristics. *Computer Music Journal*, 2004.

[5] M. Gales, D. Pye, and P. Woodland. Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation. In *ICSLP*, pages 1832–1835, 1996.

[6] 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.

[7] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds. Integrated models of signal and background with application to speaker identification in noise. *IEEE Trans. Speech and Audio*, 2(2):245–257, April 1994.

[8] J. A. Fessler and A. O. Hero. Space-alternating generalized expectation-maximization algorithm. *IEEE Trans. on Signal Processing*, 42(10), Oct. 1994.

[9] R. Gribonval, L. Benaroya, E. Vincent and C. Févotte. Proposals for performance measurement in source separation. In *ICA*, pages 763–768, April 2003.