Robust singer identification in polyphonic music using melody enhancement and uncertainty-based learning

Mathieu Lagrange, Alexey Ozerov, Emmanuel Vincent

ROBUST SINGER IDENTIFICATION IN POLYPHONIC MUSIC USING MELODY ENHANCEMENT AND UNCERTAINTY-BASED LEARNING

Mathieu Lagrange
STMS - IRCAM - CNRS - UPMC
mathieu.lagrange@ircam.fr

Alexey Ozerov
Technicolor Research & Innovation, France
alexey.ozerov@technicolor.com

Emmanuel Vincent
INRIA, Centre de Rennes - Bretagne Atlantique
emmanuel.vincent@inria.fr

ABSTRACT

Enhancing specific parts of a polyphonic music signal is believed to be a promising way of breaking the glass ceiling that most Music Information Retrieval (MIR) systems are now facing. The use of signal enhancement as a pre-processing step has led to limited improvement though, because distortions inevitably remain in the enhanced signals that may propagate to the subsequent feature extraction and classification stages. Previous studies attempting to reduce the impact of these distortions have relied on the use of feature weighting or missing feature theory. Based on advances in the field of noise-robust speech recognition, we represent the uncertainty about the enhanced signals via a Gaussian distribution instead that is subsequently propagated to the features and to the classifier. We introduce new methods to estimate the uncertainty from the signal in a fully automatic manner and to learn the classifier directly from polyphonic data. We illustrate the results by considering the task of identifying, from a given set of singers, which one is singing at a given time in a given song. Experimental results demonstrate the relevance of our approach.

1. INTRODUCTION

Being able to focus on specific parts of a polyphonic musical signal is believed to be a promising way of breaking the glass ceiling that most Music Information Retrieval (MIR) tasks are now facing [3]. Many approaches were recently proposed to enhance specific signals (e.g., vocals, drums, bass) by means of source separation methods [7, 19].

The benefit of signal enhancement has already been proven for several MIR classification tasks, such as singer identification [10, 16], instrument recognition [12], tempo estimation [4], and chord recognition [20]. In most of those works, signal enhancement was used as a pre-processing step. Since the enhancement process must operate with limited prior knowledge about the properties of the specific parts to be enhanced, distortions inevitably remain in the enhanced signals that propagate to the subsequent feature extraction and classification stages resulting in limited improvement or even degradation of the classification accuracy.

A few studies have attempted to reduce the impact of these distortions on the classification accuracy. In [10, 15], feature weighting and frame selection techniques were proposed that associate a constant reliability weight to each feature over all time frames or to all features in each time frame. In practice, however, distortions affect different features in different time frames so that the assumption of constant reliability does not hold. A more powerful approach consists of estimating and exploiting the reliability of each feature within each time frame. A first step in this direction was taken in [8], where recognition of musical instruments in polyphonic audio was achieved using the missing feature theory. This theory adopted from noise-robust speech recognition assumes that only certain features are observed in each time frame while other features are missing and thus discarded from the classification process [5].

Nevertheless, the approach in [8] has the following three limitations. First, such *binary uncertainty* (either observed or missing) does not account for partially distorted features nor for correlations between the distortions affecting different features. To avoid this limitation, it was proposed in the speech recognition field to use the so-called *Gaussian uncertainty* [6], where the distortions over a feature vector are modeled as a zero-mean multivariate Gaussian with possibly non-diagonal covariance matrix. Second, this approach necessitates clean data to train the classifiers, while for some tasks, e.g., singer identification, collecting such clean data may be impossible. Third, the approach in [8] relies on manual f0 annotation and its use in a fully automatic system has not been demonstrated.

The contribution of this paper is threefold: (1) promoting the use of Gaussian uncertainty instead of binary uncertainty for robust classification in the field of MIR, (2) using a fully automatic procedure for Gaussian uncertainty estimation, (3) learning classifiers directly from noisy data with Gaussian uncertainty.

To illustrate the potential of the proposed approach we consider in this paper the task of singer identification in popular music and address it, in line with [10, 16], using Gaussian Mixture Model (GMM)-based classifiers and Mel-frequency cepstral coefficients (MFCCs) as features. We consider this task since it is one of the MIR classification tasks for which the benefit of signal enhancement is

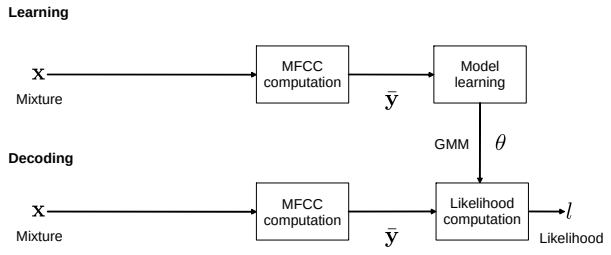

Figure 1. The standard classification scheme.

most obvious. Indeed, the information about singer identity is mostly concentrated in the singing voice signal.

The remainder of this paper is organized as follows. Some background about singer identification and baseline approaches is provided in Section 2. The proposed approach based on Gaussian uncertainty is detailed in Section 3. Experiments are presented in Section 4 and a discussion is provided in Section 5.

2. SINGER IDENTIFICATION

2.1 Background

When it comes to characterizing a song from its content, identifying the singer that is performing at a given time in a given song is arguably an interesting and useful piece of information. Indeed, most listeners have a strong commitment to the singer while listening to a given song. However, the literature about automatic singer identification is relatively scarce, compared for example with musical genre detection. This may be explained by several difficulties that pose interesting challenges for research in machine listening.

First, the human voice is a very flexible and versatile instrument and very small changes in its properties have noticable effects on human perception. Second, the musical accompaniment that forms the background is very diverse and operates at about the same loudness as the singing voice. Hence, very little can be assumed on both sides and the influence of the background cannot be neglected.

For humans, though, it is relatively easy to focus on the melody sung by the singer as our hearing system is highly skilled at segregating human vocalizations within cluttered acoustical environments. This segregation is also made possible by compositional choices. For example, most of the time in pop music, only one singer is singing at a time, and if not, the others are background vocals that are usually more easily predictable and sung at a relatively low volume.

From an application perspective, singing voice enhancement is expected to be useful for the identification of singers which have sung with different bands or with different instrumentations, such as unplugged versions. More on the so-called album effect can be found in [14]. In this case, classifying the mixture signal will induce variability in the singer models due to occlusion, while classifying the singing voice signal alone should provide better identification. The same remark applies to the case where a song features multiple singers and one needs to identify which singer is singing at a given time. For some other repertoires where the notions of singer and artist/band are very tightly linked, it is questionable whether the singing voice signal suffices for classification, because the musical background can also provide discriminative cues. Nevertheless, singing voice enhancement is likely to remain beneficial by enabling the computation of separate features over the singing voice and over the background and their fusion in the classification process. In this paper, for simplicity, we illustrate the potential of our approach by considering the singing voice signal only unless otherwise stated.

2.2 Baseline Approaches

More formally, let us assume that each recording x_{fn} (also called mixture), represented here directly in the Short Term Fourier Transform (STFT) domain, $f = 1, \ldots, F$ and $n = 1, \ldots, N$ being respectively frequency and time indices, is the sum of two contributions: the main melody (here the singing voice) v_{fn} and the accompaniment a_{fn}. This can be written in the following vector form:

$$\mathbf{x}_n = \mathbf{v}_n + \mathbf{a}_n, \tag{1}$$

where $\mathbf{x}_n = [x_{1n}, \ldots, x_{Fn}]^T$, $\mathbf{v}_n = [v_{1n}, \ldots, v_{Fn}]^T$ and $\mathbf{a}_n = [a_{1n}, \ldots, a_{Fn}]^T$.

We assume that there are K singers to be recognized, and for each singer there is a sufficient amount of training and testing mixtures. In line with [10, 16], we adopt a singer identification approach based on MFCC features and GMMs.

Without any melody enhancement such an approach consists in the following two steps [13] (Fig. 1):

1. *Learning:* For each singer $k = 1, \ldots, K$, the corresponding GMM model is estimated in the maximum likelihood (ML) sense from the features (here MFCCs) $\bar{\mathbf{y}}$ computed directly from the training mixtures of that singer.

2. *Decoding:* A testing mixture $\mathbf{x}$ is assigned to the singer k for which the likelihood of model θ_k evaluated on the features extracted in the same way is maximum [1] .

In order to gain invariance with respect to the accompaniment, one needs to separate the contribution of the accompaniment and the singer within the mixture. This separation may be embedded within the classifier, as in [22]. In this case, the separation has to be performed in the feature domain, usually the log Mel spectrum.

Alternatively, melody enhancement can be applied as a pre-processing step [10, 16] over the spectrogram of the mixture. since the spectrogram have better spectral resolution than the log Mel spectrum, this approach can potentially achieve better discrimination, as in that case, the features (MFCCs) are no longer computed from the audio mixture, but from the corresponding melody estimate $\bar{\mathbf{v}}$ (Fig. 2).

[1] In order not to overload the notations, the singer index k is omitted hereafter, where applicable.

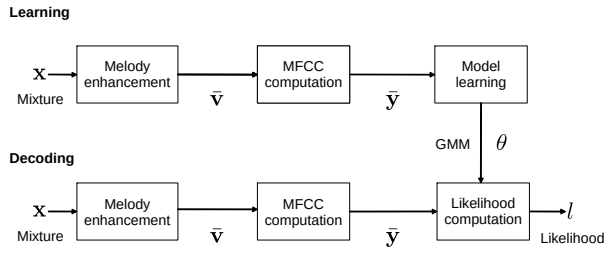

Figure 2. Considering melody enhancement as a pre-processing step.

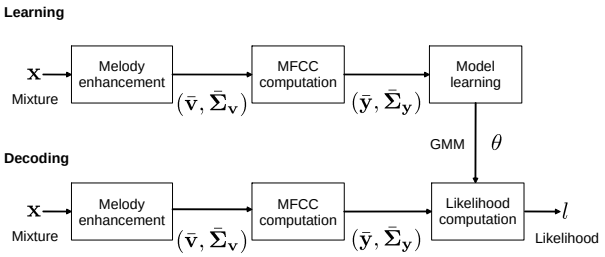

Figure 3. Proposed approach with melody enhancement and Gaussian uncertainty.

3. PROPOSED APPROACH

Inspired by some approaches in speech processing [6], we propose to consider Gaussian uncertainty by augmenting the melody estimates $\bar{\mathbf{v}}$ by a set of covariance matrices $\mathbf{\Sigma_v}$ representing the errors about these estimates. This Gaussian uncertainty is first estimated in the STFT domain, then propagated through MFCC computation, and finally exploited for GMM learning and decoding steps (Fig. 3).

3.1 Melody Enhancement

Given the mixture, we assume that each STFT frame $\mathbf{v}_n$ of the melody is distributed as

$$\mathbf{v}_n|\mathbf{x}_n \sim \mathcal{N}(\bar{\mathbf{v}}_n, \bar{\mathbf{\Sigma}}_{\mathbf{v},n}), \tag{2}$$

and we are looking for an estimate of $\bar{\mathbf{v}}_n$ and $\bar{\mathbf{\Sigma}}_{\mathbf{v},n}$.

In this study, we have chosen the melody enhancement method [2] proposed by Durrieu *et al.* [7]. This method has shown very promising results for vocals enhancement task within the 2011 Signal Separation Evaluation Campaign (SiSEC 2011) [2] and its underlying probabilistic model facilitates STFT domain uncertainty computation.

The main melody **v**, usually a singer, is modeled thanks to a source/filter model, and the accompaniment **a** is modeled using Non-negative Matrix Factorization (NMF) model. The leading voice is assumed to be harmonic and monophonic. The separation system mainly tracks the leading voice following two cues: first its energy, and second the smoothness of the melody line. Therefore, the resulting separated leading voice is usually the instrument or voice that is the most salient in the mixture, over certain durations of the signal. Overall this modeling falls into the framework of constrained hierarchical NMF with Itakura-Saito divergence [19], which allows a probabilistic Gaussian interpretation [9].

More precisely the method is designed for stereo mixtures. Let mixing equation

$$\underline{x}_{j,fn} = \underline{v}_{j,fn} + \underline{a}_{j,fn} \tag{3}$$

be a stereophonic version of the monophonic mixing equation (1), where $j = 1, 2$ is the channel index and equations (1) and (3) are related for any signal $\underline{s}_{j,fn}$ as

$$s_{fn} = (\underline{s}_{1,fn} + \underline{s}_{2,fn})/2. \tag{4}$$

[2] The Python source code is available at `http://www.durrieu.ch/research/jstsp2010.html`

A probabilistic Gaussian interpretation of modeling in [7] assumes $\underline{v}_{j,fn}$ and $\underline{a}_{j,fn}$ are zero-mean Gaussians that are mutually independent and independent over channel j, frequency f and time n. The corresponding constrained hierarchical NMF structured modeling allows the estimation of their respective variances $\sigma^2_{\underline{v},j,fn}$ and $\sigma^2_{\underline{a},j,fn}$ from the multichannel mixture. With these assumptions the posterior distribution of $\underline{v}_{j,fn}$ given $\underline{x}_{j,fn}$ can be shown to be Gaussian with mean

$$\bar{\underline{v}}_{j,fn} = \frac{\sigma^2_{\underline{v},j,fn}}{\sigma^2_{\underline{v},j,fn} + \sigma^2_{\underline{a},j,fn}} \underline{x}_{j,fn} \tag{5}$$

obtained by Wiener filtering, as in [7], and the variance [21]

$$\bar{\sigma}^2_{\underline{v},j,fn} = \frac{\sigma^2_{\underline{v},j,fn}\sigma^2_{\underline{a},j,fn}}{\sigma^2_{\underline{v},j,fn} + \sigma^2_{\underline{a},j,fn}}. \tag{6}$$

Finally, thanks to the posterior between-channel independence of $\underline{v}_{j,fn}$ and the down-mixing (4), $\bar{\mathbf{v}}_n$ and $\bar{\mathbf{\Sigma}}_{\mathbf{v},n}$ in (2) are computed as

$$\bar{\mathbf{v}}_n = \left\{(\bar{\underline{v}}_{1,fn} + \bar{\underline{v}}_{2,fn})/2\right\}_f, \tag{7}$$

$$\bar{\mathbf{\Sigma}}_{\mathbf{v},n} = \operatorname{diag}\left[\left\{(\bar{\sigma}^2_{\underline{v},1,fn} + \bar{\sigma}^2_{\underline{v},2,fn})/2\right\}_f\right]. \tag{8}$$

Note that any Gaussian model-based signal enhancement method, e.g., one of the methods implementable via the general source separation framework in [19], is suitable to compute this kind of uncertainty in the time-frequency domain.

3.2 Uncertainty Propagation during MFCC Computation

Let $\mathcal{M}(\cdot)$ be the nonlinear transform used to compute an M-dimensional MFCC feature vector $\mathbf{y}_n \in \mathbb{R}^M$. It can be expressed as [1]

$$\mathbf{y}_n = \mathcal{M}(\mathbf{v}_n) = \mathbf{D}\log(\mathbf{M}|\mathbf{v}_n|), \tag{9}$$

where $\mathbf{D}$ is the $M \times M$ DCT matrix, $\mathbf{M}$ is the $M \times F$ matrix containing the Mel filter coefficients, and $|\cdot|$ and $\log(\cdot)$ are both element-wise operations.

In line with (2), we assume that the clean (missing) feature $\mathbf{y}_n = \mathcal{M}(\mathbf{v}_n)$ is distributed as

$$\mathbf{y}_n|\mathbf{x}_n \sim \mathcal{N}(\bar{\mathbf{y}}_n, \bar{\mathbf{\Sigma}}_{\mathbf{y},n}), \tag{10}$$

which is an approximation because of the Gaussian assumption (2) and the nonlinear nature of $\mathcal{M}(\cdot)$.

To compute the feature estimate $\bar{\mathbf{y}}_n$ and its Gaussian uncertainty covariance $\bar{\mathbf{\Sigma}}_{\mathbf{y},n}$ we propose to use the Vector Taylor Series (VTS) method [17] that consists in linearizing the transform $\mathcal{M}(\cdot)$ by its first-order vector Taylor expansion in the neighborhood of the voice estimate $\bar{\mathbf{v}}_n$:

$$\mathbf{y}_n = \mathcal{M}(\mathbf{v}_n) \approx \mathcal{M}(\bar{\mathbf{v}}_n) + J_{\mathcal{M}}(\bar{\mathbf{v}}_n)\,(\mathbf{v}_n - \bar{\mathbf{v}}_n), \quad (11)$$

where $J_{\mathcal{M}}(\bar{\mathbf{v}}_n)$ is the Jacobian matrix of $\mathcal{M}(\mathbf{v}_n)$ computed in $\mathbf{v}_n = \bar{\mathbf{v}}_n$. This leads to the following estimates of the noisy feature value $\bar{\mathbf{y}}_n$ and its uncertainty covariance $\bar{\mathbf{\Sigma}}_{\mathbf{y},n}$ (10), as propagated through this (now linear) transform:

$$\bar{\mathbf{y}}_n = \mathcal{M}(\bar{\mathbf{v}}_n), \quad (12)$$

$$\bar{\mathbf{\Sigma}}_{\mathbf{y},n} = \mathbf{D}\frac{\mathbf{M}}{\mathbf{M}|\bar{\mathbf{v}}_n|\mathbf{1}_{1\times F}}\bar{\mathbf{\Sigma}}_{\mathbf{v},n}\left[\mathbf{D}\frac{\mathbf{M}}{\mathbf{M}|\bar{\mathbf{v}}_n|\mathbf{1}_{1\times F}}\right]^T, \quad (13)$$

where $\mathbf{1}_{1\times F}$ is an $1 \times F$ vector of ones and the magnitude $|\cdot|$ and the division are both element-wise operations.

3.3 GMM Decoding and Learning with Uncertainty

Each singer is modeled by a GMM $\theta = \{\boldsymbol{\mu}_i, \mathbf{\Sigma}_i, \omega_i\}_{i=1}^I$, where $i = 1, \ldots, I$ are mixture component indices, and $\boldsymbol{\mu}_i$, $\mathbf{\Sigma}_i$ and ω_i ($\sum_i \omega_i = 1$) are respectively the mean, the covariance matrix and the weight of the i-th component. In other words, each clean feature vector $\mathbf{y}_n$ is modeled as follows:

$$p(\mathbf{y}_n|\theta) = \sum\nolimits_{i=1}^{I} \omega_i N(\mathbf{y}_n|\boldsymbol{\mu}_i, \mathbf{\Sigma}_i), \quad (14)$$

where

$$N(\mathbf{y}_n|\boldsymbol{\mu}_i, \mathbf{\Sigma}_i) \triangleq \frac{1}{\sqrt{(2\pi)^M|\mathbf{\Sigma}_i|}}\left[-\frac{(\mathbf{y}_n - \boldsymbol{\mu}_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{y}_n - \boldsymbol{\mu}_i)}{2}\right]. \quad (15)$$

Since the clean feature sequence $\mathbf{y} = \{\mathbf{y}_n\}_n$ is not observed, its likelihood, given model θ, cannot be computed using (14). Thus in the "likelihood computation" step (Fig. 3), we rather compute the likelihood of the noisy features $\bar{\mathbf{y}}$ given the uncertainty and the model, that can be shown to be equal to [6]:

$$p(\bar{\mathbf{y}}|\bar{\mathbf{\Sigma}}_{\mathbf{y}}, \theta) = \prod_{n=1}^{N}\sum_{i=1}^{I}\omega_i N(\bar{\mathbf{y}}_n|\boldsymbol{\mu}_i, \mathbf{\Sigma}_i + \bar{\mathbf{\Sigma}}_{\mathbf{y},n}). \quad (16)$$

We see that in this likelihood Gaussian uncertainty covariance $\bar{\mathbf{\Sigma}}_{\mathbf{y},n}$ adds to the prior GMM covariance $\mathbf{\Sigma}_i$, thus adaptively decreasing the effect of signal distortion.

In the "model learning" step (Fig. 3), we propose to estimate the GMM parameters θ by maximizing the likelihood (16). This can be achieved via the iterative Expectation-Maximization (EM) algorithm introduced in [18] and summarized in Algorithm 1. The derivation of this algorithm is omitted here due to lack of space and the Matlab source code for GMM decoding and learning is available at `http://bass-db.gforge.inria.fr/amulet`.

Algorithm 1 One iteration of the EM algorithm for the likelihood integration-based GMM learning from noisy data.

E step. Conditional expectations of natural statistics:

$$\gamma_{i,n} \propto \omega_i N(\bar{\mathbf{y}}_n|\boldsymbol{\mu}_i, \mathbf{\Sigma}_i + \bar{\mathbf{\Sigma}}_{\mathbf{y},n}), \quad \text{and} \quad \sum\nolimits_i \gamma_{i,n} = 1, \quad (17)$$

$$\hat{\mathbf{y}}_{i,n} = \mathbf{W}_{i,n}\,(\bar{\mathbf{y}}_n - \boldsymbol{\mu}_i) + \boldsymbol{\mu}_i, \quad (18)$$

$$\widehat{\mathbf{R}}_{\mathbf{yy},i,n} = \hat{\mathbf{y}}_{i,n}\hat{\mathbf{y}}_{i,n}^T + (\mathbf{I} - \mathbf{W}_{i,n})\,\mathbf{\Sigma}_i, \quad (19)$$

where

$$\mathbf{W}_{i,n} = \mathbf{\Sigma}_i\left[\mathbf{\Sigma}_i + \bar{\mathbf{\Sigma}}_{\mathbf{y},n}\right]^{-1}. \quad (20)$$

M step. Update GMM parameters:

$$\omega_i = \frac{1}{N}\sum_{n=1}^{N}\gamma_{i,n}, \quad (21)$$

$$\boldsymbol{\mu}_i = \frac{1}{\sum_{n=1}^{N}\gamma_{i,n}}\sum_{n=1}^{N}\gamma_{i,n}\hat{\mathbf{y}}_{i,n}, \quad (22)$$

$$\mathbf{\Sigma}_i = \frac{1}{\sum_{n=1}^{N}\gamma_{i,n}}\sum_{n=1}^{N}\gamma_{i,n}\widehat{\mathbf{R}}_{\mathbf{yy},i,n} - \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T. \quad (23)$$

4. EXPERIMENTS

4.1 Database

For our evaluation, we consider a subset of the RWC Popular Music Database [11] which has previously been considered in [10] for the same task. It consists of 40 songs sung by 10 singers, five of which were male (denoted by a to e) and the five others female (denoted by f to j). This set is then divided into the four groups of songs considered in [10], each containing one song by each singer .

Each of those songs is then split into segments of 10 seconds duration. Among those segments, only the ones where a singing voice is present (not necessarily during the whole duration of the segment) are kept unless otherwise stated.

Considering short duration segments instead of the whole song is done for two reasons. First, it makes the task more generic in the sense that multiple singers can also potentially be tracked within a same song. Second, it allows us to gain statistical relevance during the cross validation by enlarging the number of tests.

4.2 Methods

For each of those segments, features are computed and classified using the three methods depicted in Figures 1 to 3. The first one, named *mix*, consists in computing the features directly from the mixture, and serves as a baseline. The second method, termed *v-sep*, consider melody enhancement as a pre-processing step. The main melody enhancement system considered in this study is available under two versions: a version focusing on the voiced part of

Accuracy (%)	per 10 sec. singing segment					per song	
input	Fold 1	Fold 2	Fold 3	Fold 4	Total	all seg.	sung seg.
mix	51	53	55	38	49	57	64
v-sep	60	63	53	43	55	57	64
v-sep-uncrt	**71**	**72**	**84**	**83**	**77**	**85**	**94**

Table 1. Average accuracy of the tested methods per singing segment and per song, considering either all the segments or only those segments where singing voice is present in the latter case.

the singing voice and another version attempting to jointly enhance the voiced and the unvoiced parts of the singing voice (see [7] for details). In the following, only the results of the former are reported since the latter led to much smaller classification accuracy. When the estimated vocals signal has zero power in a given time frame, the resulting MFCCs may be undefined. Such frames are discarded. The last method, termed *v-sep-uncrt*, consists in exploiting the estimated uncertainty about the enhancement process.

For all those methods, we considered MFCC features and dropped the first coefficient, thus discarding energy information. Mixtures of 32 Gaussians are then trained using 50 iterations of the EM algorithm for each singer. For testing, the likelihood of each singer model is computed for each segment and the one with the highest likelihood is selected as the estimate.

4.3 Results

The aforementioned four groups of songs are considered for a 4-fold cross validation. For each fold, the selected group is used for testing and the data of the three remaining ones are used for training the models. The average detection accuracy are shown in Table 1. Compared to the baseline, *v-sep* and *v-sep-uncrt* achieve better performance while considering segments, indicating that focusing on the main harmonic source within the segment is beneficial for identifying the singer. That is, the level of feature invariance gained by the separation process more than compensates for the distortions it induces.

Considering the uncertainty estimate adds a significant level of improvement in the $v - sep$ case. We assume that this gain of performance is obtained because the use of uncertainty allows us to focus on the energy within the spectrogram that effectively belongs to the voice and that the use of the uncertainty allows us to robustly consider standard features (MFCCs).

Performing a majority vote over the all the segments (in this case the likelihood of each singer is taken into account even if no singing voice is present) of each song gives an accuracy of 85% and restricting the vote to only the sung segments gives a 94% accuracy. These numbers can respectively be considered as worst and best cases. It is therefore likely that a complete system that would incorporate a music model to discard segments with only music would achieve an accuracy that is between those bounds. Although a more formal comparison would be needed, we believe that those results compare favorably with the performances obtained in [10] using specialized features on the same dataset while standard MFCC features were used here. It is also interesting to notice that in this case of song-level decisions, considering the separation without uncertainty does not give any improvement compared to the *mix* baseline.

5. DISCUSSION

We have presented in this paper a computational scheme for extracting meaningful information in order to tackle a music retrieval task: singer identification. This is done by considering an enhanced version of the main melody that is more or less reliable in specific regions of the time/frequency plane. Instead of blindly making use of this estimate, we propose in this paper to consider how uncertain the separation estimate is during the modeling phase. This allows us to give more or less importance to the features depending on how reliable they are in different time frames, both during the training and the testing phases. For that purpose, we adopted the Gaussian uncertainty framework and introduced new methods to estimate the uncertainty in a fully automatic manner and to learn GMM classifiers directly from polyphonic data.

One should notice that the proposed scheme is not tied to the task considered in this paper. It is in fact completely generic and may be easily applied to other GMM-based MIR classification tasks where the prior isolation of a specific part of the music signal could be beneficial. The only part that would require adaptation is the derivation of VTS uncertainty propagation equations for other features than MFCCs. Uncertainty handling for other classifiers than GMM has also received some interest recently in the speech processing community.

The experiments reported in this paper provide us with encouraging results. Concerning this specific task of singer identification, we intend to exploit both the enhanced singing voice and accompaniment signals and to experiment on other datasets with a wider range of musical styles. In particular, we believe that the hip-hop/rap musicals genres would be an excellent testbed both from a methodological and application point of view, as many songs feature several singers: knowing which singer is performing at a given time is a useful piece of information. Finally, we would like to consider other content based retrieval tasks in order to study the relevance of this scheme for a wider range of applications.

6. ACKNOWLEDGMENTS

This work was partly supported by the Quaero project funded by Oseo and ANR-11-JS03- 005-01. The authors wish to

thank Jean-Louis Durrieu for kindly providing his melody estimation source code to the community and Mathias Rossignol for his useful comments.

7. REFERENCES

[1] K. Adiloğlu and E. Vincent. An uncertainty estimation approach for the extraction of individual source features in multisource recordings. In *EUSIPCO, 19th European Signal Processing Conference*, 2011.

[2] S. Araki, F. Nesta, E. Vincent, Z. Koldovsky, G. Nolte, A. Ziehe, and A. Benichoux. The 2011 signal separation evaluation campaign (SiSEC2011): - audio source separation. In *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation*, pages 414–422, 2012.

[3] J.-J. Aucouturier and F Pachet. Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

[4] P. Chordia and A. Rae. Using source separation to improve tempo detection. In *Proc. 10th Intl. Society for Music Information Retrieval Conference*, pages 183–188, Kobe, Japan, 2009.

[5] M Cooke. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34(3):267–285, June 2001.

[6] 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*, 13(3):412–421, 2005.

[7] J.L. Durrieu, G. Richard, B. David, and C. Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):564–575, 2010.

[8] J. Eggink and G. J. Brown. Application of missing feature theory to the recognition of musical instruments in polyphonic audio. In *Proc. 4th International Conference on Music Information Retrieval*, 2003.

[9] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3):793–830, Mar. 2009.

[10] H. Fujihara, M. Goto, T. Kitahara, and H. G. Okuno. A modeling of singing voice robust to accompaniment sounds and its application to singer music information retrieval. *IEEE Trans. Audio, Speech and Language Processing*, 18(3):638–648, 2010.

[11] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: popular, classical, and jazz music databases. *Proc. International Conference for Music Information Retrieval (ISMIR)*, pages 287–288, 2003.

[12] T. Heittola, A. Klapuri, and T. Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. *in Proc. 10th Intl. Society for Music Information Retrieval Conference, Kobe, Japan*, pages 327–332, 2009.

[13] Y. E. Kim. Singer identification in popular music recordings using voice coding features. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Paris, France, 2002.

[14] Y. E. Kim, D. S. Williamson, and S. Pilli. Towards quantifying the album effect in artist identification. In *Proc. of Int.Conf. on Music Information Retrieval (ISMIR)*, pages 393–394, 2006.

[15] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. Instrument identification in polyphonic music: Feature weighting to minimize influence of sound overlaps. *EURASIP Journal on Advances in Signal Processing*, 2007, 2007. article ID 51979.

[16] A. Mesaros and T. Virtanen. Singer identification in polyphonic music using vocal separation and pattern recognition methods. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 375–378, 2007.

[17] P. J. Moreno, B. Raj, and R. M. Stern. A vector Taylor series approach for environment-independent speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96)*, volume 2, pages 733 – 736, 1996.

[18] A. Ozerov, M. Lagrange, and E. Vincent. GMM-based classification from noisy features. In *Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, pages 30–35, Florence, Italy, September 2011.

[19] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 20(4):1118 – 1133, 2012.

[20] J. Reed, Y. Ueda, S. M. Siniscalchi, Y. Uchiyama, S. Sagayama, and C. H. Lee. Minimum classification error training to improve isolated chord recognition. In *Proc. 10th Intl. Society for Music Information Retrieval Conference*, pages 609–614, Kobe, Japan, 2009.

[21] 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 Processing*, 2(2):245–257, April 1994.

[22] W.H. Tsai and H.M. Wang. Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. *IEEE Transactions on Audio Speech and Language Processing*, pages 1–35, 2006.