

COMBINING CLASSIFICATIONS BASED ON LOCAL AND GLOBAL FEATURES: APPLICATION TO SINGER IDENTIFICATION

Lise Regnier,

IRCAM
Paris, France

`lise.regnier@ircam.fr`

Geoffroy Peeters,

IRCAM
Paris, France

`geoffroy.peeters@ircam.fr`

ABSTRACT

In this paper we investigate the problem of singer identification on acapella recordings of isolated notes. Most of studies on singer identification describe the content of signals of singing voice with features related to the timbre (such as MFCC or LPC). These features aim to describe the behavior of frequencies at a given instant of time (local features). In this paper, we propose to describe sung tone with the temporal variations of the fundamental frequency (and its harmonics) of the note. The periodic and continuous variations of the frequency trajectories are analyzed on the whole note and the features obtained reflect expressive and intonative elements of singing such as vibrato, tremolo and portamento. The experiments, conducted on two distinct data-sets (lyric and pop-rock singers), prove that the new set of features capture a part of the singer identity. However, these features are less accurate than timbre-based features. We propose to increase the recognition rate of singer identification by combining information conveyed by local and global description of notes. The proposed method, that shows good results, can be adapted for classification problem involving a large number of classes, or to combine classifications with different levels of performance.

1. INTRODUCTION

The goal of classification is to assign unlabeled patterns into a number of known categories. A system of classification is based on an appropriate description of the patterns (features) and on a statistical algorithm (classifier) trained to learn the specificities of each pattern for the given problem. To evaluate the performance of a system, a new set of data is given to the classifier and the portion of patterns assigned to their correct class is given as an indicator of its global performance. However, each system of classification has its own limitation. To increase the classification accuracy it is necessary to introduce and combine complementary information on the problem (either new representations of the patterns or new specifications of the classes).

Classification of speech and musical signals has been largely investigated this last decade and all sort of features (temporal and spectral) and classifiers have been tested. A special attention has been given to features related to the timbre. Timbre is a perceptual attribute of the sounds that seems to be multi-dimensional. However, research has shown that timbre can be partly transcribed by the spectral envelope of sounds. In speech processing area, the source-filter model [1] clearly justify the use of spectral envelope for speech recognition problem. Researches on instruments recognition have also proved that spectral envelope was a good element to discriminate musical instruments. The singing voice is a musical instrument that has much in common with speech. For this

reason, most of works carried out on the topic of singer identification have based the description of sung signals on features derived from the spectral envelope. They have obtained satisfying results with this approach but to improve the identification performance it is necessary to extract additional information on the signals of singing voice.

In this study we suggest to describe signals of singing voice with intonative and expressive elements characteristic of singing. We propose a new set of features derived from the analysis of the trajectory of the fundamental frequency (and its harmonics). In a previous work [2] we have demonstrate that elements such as vibrato, tremolo and portamento were efficient to detect the presence of singing voice within a song. We propose now to evaluate if these features can be used to discriminate singers between them. More precisely, we evaluate if intonative features can be combined with timbre-based features to improve the performance of singer identification.

The combination of information is not a straightforward problem. In our case, the patterns to be classified are notes sung acapella. For each note, we extract timbre-based and intonative features. Timbre-based features are computed on short frames (local features) whereas information on intonation is obtained when considering the note globally (global features). As a result, the two descriptions have different sizes and cannot be compacted into a single feature-set without adding redundancy or deleting important information. The only solution is then to train two classifiers on each set of features separately and to combine their decisions afterwards. Working with only two decisions, simple voting methods cannot be applied. In addition we know from preliminary experiments, that timbre-based features provide much better results than intonative features. In general it is ticklish to improve a good performance by combining information less accurate. The proposed combination method is based on the class set reduction approach. It starts with the feature set leading to the best performance. The output of the classifier for this feature set is analyzed to deduce a restricted set of possible classes. The deduction is done by regarding the membership value (pseudo-posterior probability) for all the classes. The second set of feature is then used to perform the classification within the reduced set of classes. The membership values, for the remaining classes, returned by the two classifiers are then analyzed to take the final decision.

The paper is organized as follow: In section 2, we present some related works on singer identification and on combination of information. We present in section 3 the different elements of our method: the features, the classifiers and the combination method. The method is evaluated in section 4 on two data-sets composed of accapella recording of notes. Finally, section 5 summarizes the main results of the paper and offer some conducting remarks.

2. RELATED WORKS

2.1. On singer identification

Numerous researches have been carried out on the topic of singer identification (SID) because the voice is for many listeners the element that focuses the most their attention. Most of the methods propose to describe and recognize the content of audio signals with spectral features such as MFCC ([3], [9], [6]), LPC and their variants ([4],[5]). The features are given as input to classifier to construct a model per singer present in the data-set. To retrieve the singer of a query song, the features extracted from this song are compared to the models obtained in the previous step and the song is assigned to the class whose model is the most likely. In previous researches SVM [3], GMM [4] [5] [6] [7] [8] ANN [3] [9] have been tested for the task of SID. We consider that models obtained using features extracted from audio mixtures (i.e. voice + instruments) represent the identity of the artist (or music band) instead of the singer. To get models more representative of the singer it has been suggested in [9] and [4] to perform the classification using the segments of the song where the voice is present only (i.e they discard the purely instrumental portions of the song from the analysis). These methods still rely on features extracted from audio mixtures and it is therefore not possible to examine how much the results obtained are corrupted by the presence of instrumental background. To obtain information directly related to the voice on mixtures it has been proposed in [7], [8] to deduce a solo-singer model from a model obtained on purely instrumental parts of the song and a model obtained on the vocals (voice+instruments) portions. In most of these studies, they do not find any improvement by treating separately instrumental and vocal parts of the song. We can suppose that in some cases, the performances of systems based on spectral features obtained on mixtures directly are strongly affected by the "album" or "produced effect" [10] (all songs from the same album/producer share overall spectral characteristics). It has also been suggested to perform the classification on isolated vocals: in [5] the voice is isolated by reducing the instrumental accompaniment, in [6] the voice is re-synthesized using the components harmonically related to the fundamental frequency of the sung melody. Some studies ([6] or [11]) have compared results obtained on acapella recordings with results obtained on voice isolated from mixtures (where the mixtures were created using the same acapella recordings mixed with other instrumental tracks). They usually reported that the performances obtained on isolated vocals is much lower than the performances obtained on acapella recordings and suggest that the loss is due to the artifacts created when isolating the voice.

In this study we work in the ideal case of acapella recordings and suggest performing identification by combining local and global descriptions of the voice. So, before presenting the details of our method (features and combination rules) we review in the next paragraph some of the basic points of the information combination theory for classification problem.

2.2. On combination of decisions

Each system of classification (features+classifier) has its own limitation. To improve classification accuracy it has been proposed to combine complementary information on the same problem. The best way to obtain complementary information on a problem is probably to describe the patterns with different approaches. In some ideal cases the feature-sets given by the different descrip-

tions can be directly combined to form a unique feature-set (early fusion). In many cases, when the features have different types, ranges of values, size or different physical meanings, grouping all the features together can completely degrade the information conveyed by the features when considered independently. For this reason, it has been proposed to combine the decisions of the systems of classification instead (late fusion). In this case, each classification system works with its own feature-set. Combination of decisions can be grouped into two categories according to their architecture: parallel and sequential.

In **parallel combination** all systems involved in the combination have to classify the same data-set into the same known categories. Then, the final decision is taken by applying predefined rules on the decisions of all the classifiers. A classifier can return: a *single class*, a *list of classes* ordered in term of preference or a *membership value* for each class [12]. From the membership values we can deduce the ranked list of classes, from the list we can deduce the most likely class. An overview of methods developed for each type of outputs is presented in [13]. The more complete outputs, the more difficult to combine. Theoretically, it is not feasible to combine membership measurements obtained using different feature spaces or different type of classifiers because their respective values may not have different significations as explained in [14]. However, the methods of transformation presented in [15] can be applied to normalize the outputs.

The goal of the combination is to reach a higher accuracy than each of the individual classifications. In practice, when all classifications have equivalent accuracies, most of the basic combination rules (as *majority voting* or *sum-rule*) can reach this goal as long as the classification are not too correlated. However, when the systems to be combined show different levels of performance it is necessary to introduce knowledge on the relative performances into the combination rule. An easy way to realize such a combination is to consider the classifiers outputs as a new features and to train the combination rule (or a meta-classifier). Methods based on trained combiners generally show good results, but to avoid a lack of generalization, a very large amount of training data is necessary. Indeed, if the combiner is trained on the data-set used to learn the specificities of classes there is a large risk of over-fitting. To avoid this, the data set should be divided into three parts. The models should be learned on the first part and evaluated with the second part of the data. The results obtained should then be used to train the combiner. The global performance should be evaluated on the remaining part.

In **sequential combination** the classification systems are applied one after another using the output of the previous classifier to define a new problem for the next classifier. The final decision is generally given by the last classification (the decision-making process can be viewed as a decision tree). From these sequential methods, we retain:

- The *hierarchical* methods [16] that can be applied when the data has a taxonomy,
 - the *cascade* methods [17] where a pattern is processed by a new classifier until it is classified with a certain degree of confidence
 - and the *multi-stage* methods [18] that attempt to reduce the number of possible classes at each stage until one class remain possible.
- Sequential methods are shown to be specially adapted to solve problem involving a large number of classes and are particularly suitable for the recognition of rare event (i.e when the classes of the data set are not well balanced). With sequential classification, there is no possible backwards analysis. If the decision taken at

one step is wrong the full process will be affected.

3. DETAILS OF THE PROPOSED METHOD

We first introduce our local and global features used to describe the sound samples, then we briefly present the classifiers used next in the evaluation. Finally, we present the combination rule developed to combine the different type of features.

3.1. Sound description

Any sound can be considered as a pattern varying along two dimensions: the time and the frequency axis (as shown by the common representation of sound: the spectrogram). To obtain an accurate description of a sound we suggest to describe sounds over these two dimensions: (1) describe behavior of the frequencies at a given time (on one frame of few ms) and (2) describe the temporal variations of one frequency (or one band of frequencies) on a interval of time (segment of few sec). Features extracted at a given time of the signal and repeated along the signal will be referred as *local features* and features extracted on a longer interval of time will be referred as *global features*. Local features have focussed the most attention in all audio classification problems. The relative amplitude of frequencies, represented by the overall shape of the spectrum (the spectral envelope), has been proved to be efficient to describe and discriminate sounds in tasks of speaker and musical instruments recognition. In this study we work on classification of sung signals. Singing voice differs from speech in his musical intention and also in its production. One of the major difference is that sung sounds are most of the time voiced and sustained to allow intelligibility of the lyrics. On these sustained voiced sound, many characteristics can be obtained by analyzing the temporal variations of one frequency band on a given interval of time. These variations, intentional or not, enhance the singing voice in two points: first, these variations add expression but also they help the voice to stand out of the instrumental background. In the next paragraph we present the features used to describe the spectral envelope on this study. Next, we present the features extracted on the frequency trajectories.

3.1.1. Timbre: Local description of sound

Information related to the timbre is supposed to be conveyed by the spectral envelope. This idea comes from the description of speech sound using the source filter model by Fant [1]. In this model we suppose that the source is a periodic train of pulses (where the pitch of the produced sound is given by the distance between the pulses) modified by a filter: the vocal tract. The goal is to decorrelate the filter from the source to obtain an approximation of the transfer function of the vocal tract. The vocal tract enhances some frequencies (phenomene of extra resonance). The response of the filter is given by the global shape of the spectrum: the spectral envelope.

Many methods, with different theoretical backgrounds, have been developed to estimate and encode the spectral envelope. In our evaluation, we use three different representations of the spectral envelope: the coefficients derived from the Linear Predictive Analysis (LPC), the Mel Frequency Cepstral Coefficients (MFCC) and the Cepstral Coefficients derived from the True Envelope (TECC).

LPC and MFCC have been already used for the task of singer recognition and for we refer the reader to the works presented in

2.1 for a detailed description of these coefficients. The true envelope, introduced in [19], has been mostly used in the speech processing area. As shown in [20], this envelope estimation is more robust (especially for high pitched signals) than many other envelope estimation methods. Like the MFCC, the true envelope is estimated in the cepstral domain. This domain offers the possibility to transform the convolution of two signals into the addition of their spectra. So that, the cepstrum of a speech signal is the addition of the cepstrum of the vocal tract response and the cepstrum of the excitation signal. The real cepstrum of a discrete signal $x(n)$ is defined as the inverse Fourier transform of the log-amplitude spectrum of $x(n)$. If $X(k)$ designates the k^{th} point of the discrete Fourier transform (DFT) of $x(n)$ (with K the total number of point of the DFT), the cepstrum $C(m)$ of $x(n)$ is given by:

$$C(m) = \sum_{k=0}^{K-1} \log(|X(k)|) e^{\frac{2i\pi km}{K}} \quad (1)$$

True envelope estimation is based on iterative cepstral smoothing of the log-amplitude spectrum. We denote $C_i(k)$ the cepstral representation of the envelope at the i^{th} iteration for the bin k of the DFT (1). The algorithm iteratively updates the smoothed input spectrum $A_i(k)$ using the maximum of the original spectrum $|X(k)|$ and the current spectral representation.

$$A_i(k) = \max(\log(|X(k)|), C_{i-1}(k)) \quad (2)$$

The cepstral smoothing is then applied to $A_i(k)$ to obtain $C_i(k)$. The iterative algorithm stops if for all bin k and a fixed threshold τ the relation $A_i(k) < C_i(k) + \tau$ is satisfied.

At the end of this operation, the true envelope has the same size than the cepstrum. To concentrate the information conveyed by this envelop into a smaller number of coefficients, the Discret Cosine Transform (DCT) is computed on the envelop and the firsts coefficients are retained. (This method is similar to the method applied to obtained the MFCC). We named in the following, the coefficients obtained TECC.

The goal of any envelope estimation is to retain from the signal the contribution of the filter by discarding information of the pitch. For high pitched signals this estimation can be problematic since the envelope can start following the peaks related to the pitch instead of the global shape. In general, if the order of the model is low (small number of coefficients) this problem is avoided but a too low order is not be sufficient to preserve the global shape. Finding the optimal order is not straightforward. Experimentally we chose: 25 TECC, 20 MFCC and 15 LPC to model the envelope on pseudo-stationary sung signals.

3.1.2. Intonation: Global description of a note

As explained above, the singing voice differs from the speech in its musical intention and its production. Most of the sung sounds are voiced and sustained. Because of the mode of production of voice two kinds of variations appear on sustained tones:

Vibrato refers to a periodic modulation of frequency. It is a natural effect of singing voice that can be voluntary enhanced by the singer but exists naturally due to the mechanism of the singing production. When a sung tone is emitted with a vibrato, a range of frequencies (centered on the note frequency) is browsed by the vocal tract. Because of its shape, the vocal tract enhances the resonance of some frequencies. So that, depending on the morphology

of the singer, a frequency modulation is accompanied with a modulation of amplitude. The latter is referred as **tremolo** in music. In addition, when two distinct (and distant) notes are sung in the same breath, the singers pass from one to another by a continuous variation of the frequency. If the interval between the notes is small (smaller than a 3^{rd}) the smooth transition is referred to as **legato**. When the gap is higher, the transition is named **portamento**. Portamento is a singing specific term, when string instruments glides continuously from one pitch to another the transition is named glissando.

To obtained information related to periodic and continuous (resp. vibrato and portamento) of a given note, we propose to parameterize the fundamental frequency of a note with the following model:

$$f(t) = \bar{f} \cdot (d_f(t) + x(t)) + \epsilon(t) \quad (3)$$

Where \bar{f} is the mean of $f(t)$ representing the perceived pitch, $x(t)$ is a periodic modulation of frequency representing the vibrato, and $d_f(t)$ is a continuous variation of the pitch representing the portamento.

The parameters can be computed as follow:

- First, \bar{f} is given by the mean of $f(t)$. In order to get equivalent values for two partials with frequencies harmonically related ($f_p(t) = k \cdot f_q(t)$, $k \in \mathbb{Q}$), the other parameters of (3) are computed on a normalized version of the frequency trajectory: $f(t)/\bar{f}$.
- The quantity $f(t)/\bar{f}$ is low-pass filtered with a cutoff frequency $f_c = 4\text{Hz}$. The result of the filtering process is a curve representing the **relative frequency variation** $d_f(t)$ parameterized with a third-order polynom P_{df} .
- The periodic component $x(t)$ is obtained by subtracting the relative frequency deviation $d_f(t)$ from the relative frequency: $x(t) = f(t)/\bar{f} - d_f(t)$.

The vibrato term $x(n)$ can be written as a periodic modulation characterized by an amplitude (or extent) E , a frequency of modulation (or rate) r and a phase at the origin ϕ_0 :

$$x(t) = E \cdot \cos(2\pi r t + \phi_0) \quad (4)$$

The vibrato parameters (only E and r are of interest) are estimated using classical methods for sinusoidal parameters estimation.

As mentioned earlier, in singing the presence of vibrato implies the presence of an amplitude modulation (tremolo) and we suppose the relation between the two modulation singer-specific. The AM's parameters (amplitude and rate) are estimated using equations (3) and (4) applied on the function of amplitude $a(t)$.

$$a(t) = \bar{a} \cdot (d_a(t) + x(t)) + \epsilon(t) \quad (5)$$

In this case, the term \bar{a} is related to the global loudness (or the dynamic $\{p, mf, f, \dots\}$). The low variation $d_a(t)$ transcribes a possible variation of dynamic (*crescendo* for example). The sinusoidal part $x(t)$ represents the amplitude modulation itself.

In practice, the sinusoidal components (partials) are tracked along the studied sound and the analysis is performed on each partial.

3.1.3. Duality of descriptions

We can resume the features obtained on one note, using these two complementary descriptions as presented in Table 1.

Feature type	Local (see sec.3.1.1)	Global (sec.3.1.2)
Nature of the description:	Local variations (High frequencies)	Overall structural information (Low frequencies)
Analyze performed on:	p stationary portions of the signal (frames)	the whole note
Size of the description:	1 feature matrix per note $X = [\bar{x}_1, \dots, \bar{x}_p]$ where \bar{x}_i (with n coeff) is the feature vector obtained on frame i	1 feature vector per trajectory of frequency analyzed: \vec{x} , either the fundamental f_0 or p' partials analyzed
Dimension	$n \times p_{frame}$	$p'_{partials} \times n'$

Table 1: Global and local features extracted on a sung tone

3.2. Learning the specificities of classes

It exists numerous statistical algorithms for pattern recognition. We present in table 2 three classes of algorithm that differ in their approach. All of them perform supervised classification. Next, in the evaluation, we used one algorithm (the one given in example) from each class.

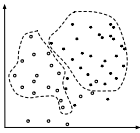
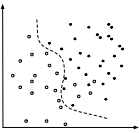
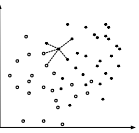
Type	Generative	Discriminative	Instance-based
General Idea			
Principle	Build one model per class	Learn the boundaries between the classes	Compare items
Example	Gaussian Mixture Model GMM	Support Vector Machine SVM	k-Nearest Neighbors kNN
Classify new pattern	Likelihood for each class	Affinity to the margins	Distance to the closest neighbors
Output	Posterior probability	Distance	Distance

Table 2: Different approaches to classify patterns

As shown on the last row of table 2 the outputs of classifiers can have different type and range in different intervals. We can consider, without lost of generality, that all the outputs have values in the interval $[0, 1]$ and represent a (pseudo) posterior probability that an item belong to one class. The transformation of classifier outputs into pseudo-posterior probability can be done using the *softmax* method proposed in [21].

Using the features and the classifier presented above we suggest to identify singer by the two types of information. In the next section we detail the method use for the combination.

3.3. Combination method

The proposed approach is a multi-stage classification method that reduces at each step the set of possible classes until a reduced set of classes remains possible. Then the membership measurements of all classifiers for each remaining class are analyzed to take a final decision.

The approach proposed here is especially adapted to:

- Combine classifications with different levels of performance.
- Solve problems involving a large number of class with no hierarchical organization of the data.
- Combine a low number of representations (when a cascade classification can not be processed until only a single class remain possible)

We first introduce the notation and then present the framework and discuss the choice of the parameters of the method that will later be applied to combine local and global features.

3.3.1. Notations

- Each pattern z (in our case one note) is assigned to one of the N possible **classes**: $\Omega = \{\omega_1 \dots \omega_N\}$.
- Each pattern can be described using different set of features, the set of available **descriptions** D_i is denoted by $\mathcal{D} = \{D_1 \dots D_L\}$.
- The set of **classifiers** is denoted by $\mathcal{C} = \{C_1 \dots C_M\}$.
- A **system of classification** is composed of one description and one classifier: $S^{(k)} = (D^{(k)}, C^{(k)})$ where $D^{(k)} \in \mathcal{D}$ and $C^{(k)} \in \mathcal{C}$.
- In our problem, the first part of the combination is done using a sequential scheme. For a given pattern z , at each step of the classification, the number of possible classes is reduced. So that, each system works with a specific $\Omega^{(k)}(z)$. If $S^{(k)}(z)$ is performed before $S^{(k+1)}(z)$, thus $\Omega^{(k+1)}(z) \subset \Omega^{(k)}(z) \subset \Omega^{(0)} = \Omega$.
- For a given classification task, all systems of classification are trained using the same data set. Since the pattern representation $D^{(k)}$ of this data-set changes for each k , the **training set** associated with $S^{(k)}$ is denoted by $T^{(k)}$. Finally, we denote by $T_{\downarrow}^{(k)}(z)$ the **training-set reduced to patterns with labels in** $\Omega^{(k)}(z)$.
- In the rest of this section we work with classifiers returning a membership measurement for each class of the problem. The output of such a classifier is denoted $C^{(k)}(z) = M^{(k)}(z) = [m_1^{(k)}(z), \dots, m_N^{(k)}(z)]$
- Working on the combination of classifier outputs we store the decision of the K classifiers for the N given classes in a decision profile matrix (of size $N \times K$) denoted by M

3.3.2. General framework

The idea behind our method is the following: The probability to retrieve the correct class of an unknown pattern increases when the number of possible classes of a given classifier decreases. So that, a classification system with a relative low accuracy can enhance the performance of a system with higher accuracy if the problem

given to the weaker system is simplified by the more accurate system. By “simplified problem” we mean a problem with a smaller number of classes.

The general framework can be summarized as follow:

The algorithm starts with a set of N class $\Omega^{(0)} = \Omega$, two descriptions of the same data set $T^{(1)}$ and $T^{(2)}$ and two classifiers $C^{(1)}$ and $C^{(2)}$. For each pattern z , system $S^{(1)}$ returns a measurement vector $M^{(1)}(z)$. The $N^{(1)}$ most likely classes are retained to form a new class-set $\Omega^{(1)} \subset \Omega^{(0)}$. The training-set used by the second system $S^{(2)}$ is reduced to patterns with classes in $\Omega^{(1)}$. The training-set derived from $T^{(2)}$ is denoted $T_{\downarrow}^{(2)}$. Then, the second classification system $S^{(2)}$ trained on $T_{\downarrow}^{(2)}$ is applied to the unknown pattern z . The process can be iterated as long as:

- The last classifier does not return a single class.
- Another system (either a new description of a new classifier) is still available.

If the method is iterated until only one class remain possible ($N^{(K)} = 1$) then the method works as a decision tree. If the process is stopped when $N^{(K)} > 1$ classes remain possible, then the rule for parallel combination can be applied on the output of the classifiers for the $N^{(K)}$ remaining classes. In opposition, if the class-set is not reduced at each step ($N^{(K)} = N$), then the method is equivalent to a classical parallel scheme of combination.

The method is illustrated in Figure 1.

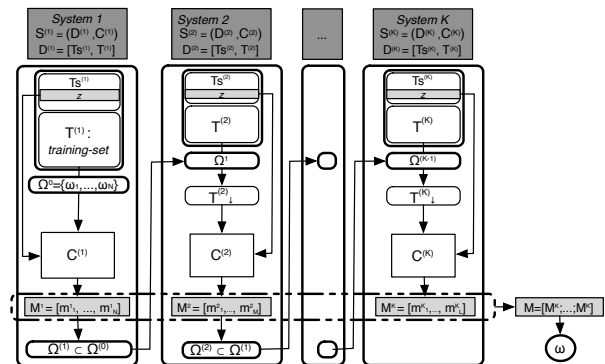


Figure 1: Scheme of the proposed method to combine K systems

We know discuss the choice of the different parameters of the method.

3.3.3. Choice of parameters

Choice of $\Omega^{(k)}$: The number can be simply defined by a relation of type: $N^{(k)} = \frac{N^{(k-1)}}{c^{(k)}}$ where $c^{(k)}$ are defined beforehand. Dynamic rules, as Bayesian information criterion or elbow method, applied on the measurements $m_n^{(k)}$, can be used to define the number of class selected at each step.

Choice of feature space and classifier: There is no restriction on the choice of the descriptions $D^{(k)}$ and the classifiers $S^{(k)}$. Thus for $i \neq j$, the combination system can be set up with $D^{(i)} = D_j$ or $C_i = C_j$. From our experiments the proposed method is still accurate if $S_i = S_j$.

Combination rule: At the end of the sequential stage $N^{(K)}$ classes remain possible. The K vectors of measurements $M^{(k)}(z)$ are reduced to values of classes in $\Omega^{(K)}$ before being combined

into a decision profile matrix M . We suggest to first normalize each column of M on the $N^{(K)}$ remaining classes, and then to apply a *sum-rule* for the reasons explained in [14].

Sequential organization of the $S^{(k)}$: If knowledge on the performances of the K systems is available, we recommend to put the best systems at the top of the iterative process. If all systems have equivalent performances, or if the relative performances cannot be estimated, the systems that require the lowest number of computation should be placed at the end of the process to reduce the cost.

4. EVALUATION OF THE PROPOSED METHOD ON A TASK OF SINGER IDENTIFICATION

We evaluate the proposed method on two distinct sets of data. Both sets are made of isolated notes and we report for each configuration tested the percentage of note assigned to their correct singer. The task is referred to as “closed-set identification” problem (i.e each note belong to one and only one singer of the set).

4.1. Data-set

The two sets are chosen for their complementarities.

The first data set, **LYR**, is composed of recordings made by 17 lyric female singers in laboratory conditions. The full description of this set is given in [22]. For each singer, the same set of tones is available (3 pitches: A5, D5, G4 sung with 3 levels of intensity: p , mf , f and each couple [pitch, intensity] is repeated 3 times). On this set, the task is referred as “closed-set, note-dependent identification” (as text dependent identification).

The second data set, **POP**, has been created by segmenting the vocal track of “pop-rock” songs into sustained notes. For each singer, we work with notes extracted from 3 songs. The task there is referred to as “close-set, note-independent” identification because each singer has a set of notes related to its tessitura. In this set, male and female singers are present.

The two data-sets can be summarized as shown in table 3.

Data-set Name	LYR	POP
Type of voice	Lyric female singers	Rock-Pop singers
Nb of singer	17 (females: F)	18 (8 Males / 10 F)
Nb of sample per singer	27 notes per singer	3 songs per singer segmented into \approx 50 notes each
Nb of sample per set	$27 \times 17 = 459$ notes	2492 notes
Recordings	Laboratory condition	Personal recording system
Nature	Isolated notes	Notes extracted from songs (in context)
Task	Note-dependent	Note-independent

Table 3: Description of the two data-sets used for the evaluation

4.1.1. Composition of training and testing set

The evaluation is done using supervised machine learning method. Both data-sets are divided into three folds: the training phase is done on the data of 2 folds and the validation is conducted on the

remaining data. Evaluation is done using a 3 folds cross-validation obtained by rotating folds, and for each experiments we report the average accuracy of the 3 experiences.

On LYR, the set of sample available for one singer can be summarized as shown in table 4. To cover the variability of one singer

LYR	p			mf			f		
A5	1	2	3	1	2	3	1	2	3
D5	1	2	3	1	2	3	1	2	3
G4	1	2	3	1	2	3	1	2	3

Table 4: Data available for one singer in LYR

and build more general models, we put into one fold data with all available pitches and intensities. To avoid having too similar data in the training and testing data-set all repetitions of the same note (pitch, intensity) are putted into the same fold. We illustrate in table 4 the repartition of the samples from singer into the 3 folds (where each color represent one fold).

On the POP data-set, each singer uses its own system of recordings and sometimes this system changes from one song to another. To ensure that the identification is performed on the singer identity and not on the song (album effect) we chose to put in one fold all notes extracted from one song. Thus, for each fold evaluation, the singer identity is learned using the notes obtained on two songs and the model obtained is tested on the notes of the remaining song.

4.2. Application of the proposed method

We now evaluate how the singer of a given note is retrieved when using local and global features independently and how the identification is enhanced when local and global features are combined with the method presented in 3.

The combination method is applied for $K=2$ systems of classification where the first system is based on local features and the second one on global features. We thus have:

• **Systems:** $S^{(1)} = (D^{(1)}, C^{(1)})$ and $S^{(2)} = (D^{(2)}, C^{(2)})$

• **Descriptions:** $\mathcal{D} = \{D_1, \dots, D_4\}$ with D_i for $i = 1 \dots 3$ are representations of the data based on local features: ($D_1 \leftarrow$ TECC, $D_2 \leftarrow$ MFCC and $D_3 \leftarrow$ LPC) and D_4 is based on global features ($D_4 \leftarrow$ INTO). Thus we have

$$D^{(1)} = D_i \text{ with } i = \{1, 2, 3\} \text{ and } D^{(2)} = D_4$$

. Experimentally we use 25 TECC, 20 MFCC, and 15 LPC.

• **Classifiers:** The available set of classifier is denoted by \mathcal{C} where $\mathcal{C} = \{C_1, C_2, C_3\} = \{SVM, GMM, SVM\}$. All possible configurations are tested for the combination:

$$\forall j, C^{(j)} = C_i \text{ with } i = 1, 2, 3$$

• **Class-set reduction rule:** The number of classes remaining at the end of the first stage is defined dynamically. The membership values are normalized such that their sum is equal to one. The classes that explain 80% of the posterior probabilities are retained to form the new subset of classes of size $N^{(1)}$.

• **Combination rule:** Once the membership measurements for the $N^{(1)}$ remaining classes have been normalized and concatenated to form a decision profile matrix, we apply a “sum-rule” to take the final decision for the reasons explained in [14].

Feature I		TECC			MFCC			LPC		
Feature II		SVM (84.97)	kNN (83.22)	GMM (77.56)	SVM (73.42)	kNN (72.55)	GMM (65.36)	SVM (80.61)	kNN (77.78)	GMM (69.06)
Into	SVM (42.48)	89.11	87.8	84.97	79.3	78.21	76.69	86.93	84.97	79.3
	GMM (42.70)	86.93	86.27	84.97	79.52	80.39	76.69	86.93	83.88	79.74
	kNN (39.22)	87.58	87.8	81.7	78.21	77.12	73.86	84.97	83.88	74.29

Table 5: Results of combination method for singer lyric singer identification (LYR)

Feature I		TECC			MFCC			LPC		
Feature II		SVM (73.57)	kNN (69.02)	GMM (69.60)	SVM (64.66)	kNN (59.26)	GMM (56.21)	SVM (69.10)	kNN (63.81)	GMM (56.17)
Into	SVM (50.12)	78.97	72.92	74.07	71.37	67.94	56.05	74.85	69.60	66.05
	GMM (46.91)	70.72	68.29	72.80	64.97	61.00	60.03	69.92	66.74	61.23
	kNN (43.25)	73.92	69.80	71.99	67.01	63.70	61.81	70.41	65.97	64.37

Table 6: Results of combination method for singer pop-rock singer identification (POP)

4.3. Results

We present in table 5 the results obtained in the LYR data-set and in table 6 the results obtained on POP data-set.

For both tables, the different configurations of $S^{(1)}$ are presented in the first row and the configurations of $S^{(2)}$ in the first column. The number into bracket placed beside the name of each classifier indicates the accuracy of the system when a single classification is applied. Finally, the accuracies of the combined classifications are reported at the intersection of the two systems used.

The task evaluated here is challenging since only a short segment (a note of few seconds length) is used to recognize the singer. We comment first results with a single type of feature and then comment the results obtained with the combination.

4.3.1. Results on single classification

Single classification with timbre-based features

From the results obtained with TECC, MFCC and LPC (first row of each table) we can deduce that timbre-based features are rather appropriate to describe voice on acapella recordings. However, we remark that results obtained on LYR are much better than results obtained with POP. In LYR, all samples have been recorded in the same ideal conditions (same mic, room) so that we can ensure that the spectral envelopes of these sounds are clearly conveying information on the vocal tract of singers. Probably the results on POP are affected by the "album-effect". We have also evaluated the performance of classification on POP when the singer models are learned on 2 thirds of each song and the validation is done on the remaining data. With a 3 folds cross-validation the average accuracy obtained is equal to 96%.

For all experiments, the TECC outperform the MFCC and LPC. In both cases, the best result is obtained when working with TECC and SVM. SVM seems to perform better than other classifiers. Even if it is not possible to ensure that each system has been optimized (transformation of the feature space, choice of the classifier parameters, etc.) we see from all these experiments that it is not possible to retrieve the singer with these features even when the task is done on acapella recordings.

Single classification with intonation-based features

Intonative features have not been yet used to singer or instrument recognition. These features can somehow find an equivalent in the prosodic features used in speaker identification. On both data-sets the classifications obtained with INTO features show a relatively good accuracy. We remind that a random classification would

have an accuracy $\approx 5\%$ when working with 17 or 18 singers. The better results obtained on POP can be easily explained. All singers in LYR have a pretty similar technique, and all their vibratos look and sound pretty similar. We do not have any information on the technique of the singers in POP but by comparing the spectrograms (and the partials) of POP singers we can see that the variety in vibrato technique is much larger. The vibrato of lyric singers generally has a large extent and it very regular but the vibrato is definitely present in pop-rock type of voices. From this experiments, we can conclude that expressive elements such as vibrato, tremolo and portamento are singer-specific. Contrary to timbre-based features, intonative features should not be affected by the "album-effect". Theoretically, the correlation between the amplitude and the frequency modulation should remain constant across the different songs of the singer. According to the analysis done on vibrato rate we can also ensure that the rate of singers' vibrato does not vary much between different songs of a singer (even if the songs have different tempo or mood).

The capacity to discriminate the classes of each feature composing INTO have been studied using the IRMSFP algorithm proposed in [16]. On the two data-sets the vibrato rate, the tremolo rate and the vibrato extent are the most discriminative features. In POP, the coefficients of the polynomial, representing the portamento, are also of importance. This is mainly due to the fact that the notes composing POP have been extracted from full song, and in many cases the segment analyzed contain note transition.

4.3.2. Results of the combination

In most of cases, combining local and global features with the proposed approach increases the identification accuracy. In average (over all experiments per data-set), a gain of 6.23% and 4.48% is obtained on LYR and POP respectively. In practice, the higher the accuracy of one system is, the more difficult will be to improve the performance by combining a system with a lower accuracy. The gap between the different systems performance is greatly reduced with the double classification. For example, if we consider the results on LYR obtained with a single classification based on LPC ($S^{(1)}, D^{(1)} = \text{LPC}$), the variance of the results obtained with any classifiers is equal to 15 ($\forall C_i, \sigma(\text{Acc}(S^{(1)})) = 15$). When the classifications based on LPC are combined with INTO features, the variance of the results is reduced to 3.44 ($\forall S^{(2)}, \sigma(\text{Acc}(S^{(1)} \cap S^{(2)})) = 3.44$).

This method of combination has been developed because no one of the traditional methods provides an amelioration of the performance already obtained with timbre-based features.

5. CONCLUSION

In this paper we have proposed a new method to identify singer using isolated notes. In the proposed method, the notes to be classified are described using local and global features representing respectively the spectral envelope and some expressive attributes specific to singing. Local descriptors such as MFCC and LPC have been previously used for this kind of experiment but we suggest to transcribe the spectral envelope with a new set of coefficients derived from the true envelope. This new set of coefficients (TECC) show better performances than coefficients traditionally used to transcribe timbre, at least for this task. In general, especially in very clean signals of singing, they perform good classification. In the case studied especially when the classes are learned with SVM. In addition, the set of global features (INTO), previously used to detect the presence of voice within songs, have been proved to be useful to characterize singer identity. They do not obtained results as good as results of timbre-based features but they have the real advantage of being completely orthogonal to the latter. In practice, it is not straightforward to find improve a good classification by introducing information yielding to a poorer classification performance. We have proposed a methods based on the idea that a system of classification with a relative low accuracy can be employed to enhance the classification returned by a stronger system if the problem given to the weaker system is simplified by the best of the two systems. In other words, for a given note, the best system is asked to deduce a subset of possible classes (as small as possible and which still contains the true class) then the second system is asked to perform the classification on the reduced set of classes. Finally, the membership measurements of the reduced set of classes returned by the two classifiers are analyzed to take the final decision. This combination method appear to be efficient for this task since the results obtained by this combination are always better than the results obtained using a single classification.

6. ACKNOWLEDGMENTS

This work was partly supported by the “Quaero” Program funded by Oseo French agency for innovation and by the bourse Dorety Leet/AFFDU. We like to thank C. Dromey for giving us the LYR data-set.

7. REFERENCES

- [1] G. Fant, “The source filter concept in voice production,” in *IV FASE Symposium on Acoustics and Speech, Venezia*, 1981.
- [2] L. Regnier and G. Peeters, “Partial clustering using a time-varying frequency model for singing voice detection,” in *ICASSP. IEEE*, 2010, pp. 441–444.
- [3] B. Whitman, G. Flake, and S. Lawrence, “Artist detection in music with minnowmatch,” *Neural Networks for Signal Processing XI, 2001. Proceedings of the 2001 IEEE Signal Processing Society Workshop*, pp. 559–568, 2001.
- [4] Y.E. Kim and B. Whitman, “Singer identification in popular music recordings using voice coding features,” in *ISMIR*, 2002, pp. 164–169.
- [5] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H.G. Okuno, “Singer identification based on accompaniment sound reduction and reliable frame selection,” *Proc. ISMIR*, pp. 329–336, 2005.
- [6] A. Mesaros, T. Virtanen, and A. Klapuri, “Singer identification in polyphonic music using vocal separation and pattern recognition methods,” in *ISMIR*, 2007, pp. 375–378.
- [7] NC Maddage, C. Xu, and Y. Wang, “Singer identification based on vocal and instrumental models,” *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 2, 2004.
- [8] W.H. Tsai and H.M. Wang, “Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals,” *Audio, Speech and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 330–341, 2006.
- [9] A. Berenzweig, D.P.W. Ellis, and S. Lawrence, “Using voice segments to improve artist classification of music,” *AES 22nd International Conference*, 2002.
- [10] Y.E. Kim, D.S. Williamson, and S. Pilli, “Towards quantifying the album effect in artist identification,” in *ISMIR*. Citeseer, 2006, pp. 393–394.
- [11] M.A. Bartsch, *Automatic singer identification in polyphonic music*, Ph.D. thesis, The University of Michigan, 2004.
- [12] L. Xu, A. Krzyzak, and C.Y. Suen, “Methods of combining multiple classifiers and their applications to handwriting recognition,” *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 22, no. 3, pp. 418–435, 2002.
- [13] L.I. Kuncheva, *Combining pattern classifiers: methods and algorithms*, Wiley-Interscience, 2004.
- [14] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, “On combining classifiers,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 3, pp. 226–239, 2002.
- [15] C.L. Liu, “Classifier combination based on confidence transformation,” *Pattern Recognition*, vol. 38, no. 1, pp. 11–28, 2005.
- [16] G. Peeters, “Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization,” *115th AES convention, New York, USA, October*, 2003.
- [17] P. Zhang, T.D. Bui, and C.Y. Suen, “A novel cascade ensemble classifier system with a high recognition performance on handwritten digits,” *Pattern Recognition*, vol. 40, no. 12, pp. 3415–3429, 2007.
- [18] T.E. Senator, “Multi-stage classification,” in *Data Mining, Fifth IEEE International Conference on*. IEEE, 2005, p. 8.
- [19] S. Imai and Y. Abe, “Spectral envelope extraction by improved cepstral method,” *Electron. and Commun. in Japan*, vol. 62, pp. 10–17, 1979.
- [20] A. Röbel, “Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation,” in *DAFx*, 2005.
- [21] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern classification*, vol. 2, Wiley, 2001.
- [22] C Dromey, N Carter, and A Hopkin, “Vibrato rate adjustment,” *Journal of Voice*, vol. 17, no. 2, pp. 168–178, 2003.