Automatic Singer Identification For Improvisational Styles Based On Vibrato, Timbre And Statistical Performance Descriptors

Nadine Kroher
Music Technology Group
Universitat Pomeu Fabra
nadine.kroher@upf.edu

Emilia Gómez
Music Technology Group
Universitat Pomeu Fabra
emilia.gomez@upf.edu

ABSTRACT

Automatically detecting the singer by analyzing audio is a challenging task which gains in complexity for polyphonic material. Related approaches in the context of Western commercial music use machine learning models which mainly rely on low-level timbre descriptors. Such systems are prone to misclassifications when spectral distortions are present, since the timbre of the singer cannot be accurately modeled. For improvisational styles, where the performance is strongly determined by spontaneous interpretation characteristic for the singer, a more robust system can be achieved by additionally modeling the singer's typical performance style. In addition to timbre and vibrato descriptors we therefore extract high-level features related to the performance character from the predominant fundamental frequency envelope and automatic symbolic transcriptions. In a case study on flamenco singing, we observe an increase in accuracy for monophonic performances when classifying on this combined feature set. We furthermore compare the performance of the proposed approach for opera singing and investigate the influence of the album effect.

1. INTRODUCTION

Music information retrieval is of great importance for managing large music databases and providing users with recommendations and automatically generated meta-data. Automatic singer identification is a complex task specially for low-quality recordings or the presence of instrumental accompaniment. Related approaches have mainly focused on the extraction of one or more timbre-related spectral descriptors such as Mel-frequency cepstral coefficients ([1], [2]; [3]), linear prediction coefficients [4] or Gamma-tone cepstral coefficients [1]. However, identification solely based on timbre description requires high quality audio recordings and is prone to errors when spectral distortions are present. More robust approaches rely on additional extraction of non-timbre features: Fujihara and Goto [5] improve the identification accuracy by additionally extracting fundamental frequency trajectories and Nwe et al. [6] use cascaded bandpass filters to estimate vibrato related descriptors. In a first genre-specific approach, Sridhar et al. [7] report an increased performance for Carnatic Music by adjusting cepstral descriptors to the intervalic structure of this particular musical material.

Despite the great variety of music traditions and corresponding communities of enthusiasts, most music information retrieval algorithms are designed and tested on databases containing mainly Western commercial music. Flamenco is a rich improvisational music tradition from Southern Spain, characterized by deviations from the Western tonal system, freedom in rhythmic interpretation and a large amount of microtonal melodic ornamentation. As an oral tradition, songs have been passed on from generation to generation and performances are often spontaneous and highly improvisational. Consequently, only rare manual scores and annotations exist. In order to design a robust singer identification system for monophonic flamenco solo singing styles, where audio recordings often lack quality, we extend a timbre-based approach by extracting vibrato descriptors from estimated fundamental frequency contours. Exploiting the improvisational character of these styles, we incorporate statistical performance descriptors obtained from automatic transcriptions. We analyze the performance for monophonic flamenco singing and explore the suitability of this approach for polyphonic flamenco and opera singing collections. Despite the limitation of this paper to a case study on Flamenco, we see a potential s improvisational singing style and furthermore consider the proposed performance descriptors for related tasks, such as performance and artist similarity characterization.

The rest of the paper is organized as follows: Section II describes a preliminary experiment and gives an oveview of feature extraction and classification of the proposed system. In section III, we present the conducted experiments and results and discuss the influence of the album effect. The conclusions of our study are summarized in section IV.

2. PROPOSED APPROACH

As it will be illustrated in the first subsection, automatic singer identification algorithms based on timbre information produce high accuracies for data sets containing high-quality audio recordings, which can not be generalized for the data investigated in this paper. Since flamenco singing is a highly improvisational art-form, a performance is influenced by a set of given rules for the style as well as the performer's individual spontaneous interpretation. We therefore explore the use of this genre-specific properties to identify a singer by modeling his or her characteristic way of performing, regarding ornamentation, dynamics, pitch content and tendencies of rhythmic and melodic interpretation. The extensive use of vocal vibrato is a key feature in flamenco singing and as in other genres, singers tend to develop particular vibrato characteristics. We therefore incorporate global vibrato descriptors as attributes in the learning task.

2.1 Baseline: Timbre-based singer identification

In a preliminary experiment we implemented a related state of the art singer identification system [2] based on framewise extracted Mel-frequency cepstral coefficients (MFCCs), as shown in figure 1.

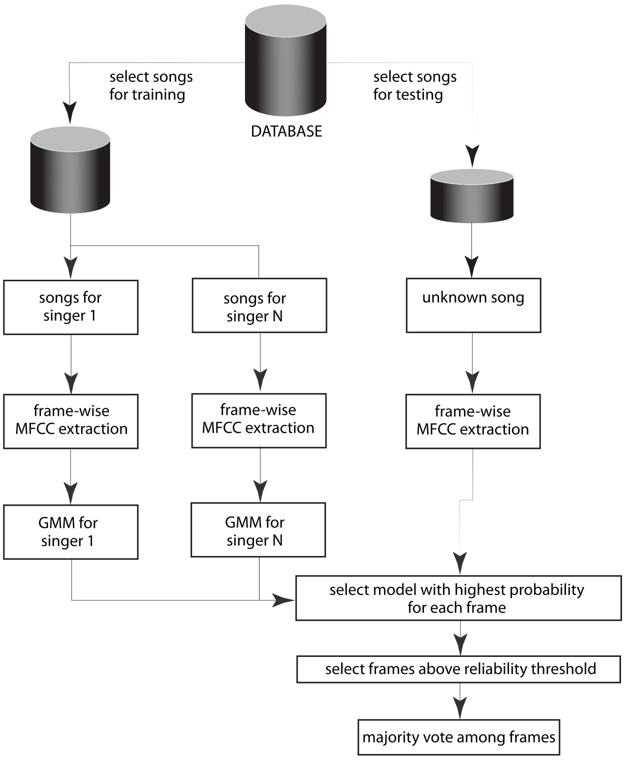
select songs for training
DATABASE
select songs for testing
songs for singer 1
songs for singer N
unknown song
frame-wise MFCC extraction
frame-wise MFCC extraction
frame-wise MFCC extraction
GMM for singer 1
GMM for singer N
select model with highest probability for each frame
select frames above reliability threshold
majority vote among frames

Figure 1. Baseline approach: Singer identification based on timbre descriptors.

We evaluated the performance for *a monophonic flamenco* singing collection containing 65 recordings as well as a subset of the *MIR-1K* database [8] containing 228 short voice only Pop music clips without accompaniment. We selected this subset in order to obtain five equally distributed classes in both sets. While the audio quality of the recordings in the *MIR-1K* is throughout high, the *monophonic flamenco* collection contains tracks with varying audio quality and includes historical recordings with strong spectral distortions. For further implementation details and database description, we refer to [9]. The system gives convincing results for the *MIR-1K* subset given the low amount of audio material (approx. 28 minutes) and confirms the suitability of a timbre-based identification, even for small databases as illustrated in table 1. By contrast, the accuracy is lower for the monophonic flamenco dataset, even though more audio material is available (approx. 2 hours 51 minutes). This lack of robustness towards varying audio quality serves as motivation to explore classification based on a fusion of timbre and non-timbre descriptors.

Database	**Correctly classified instances (CCI)**
MIR-1K subset	97.14%
Monophonic Flamenco	65.00%

Table 1. Classification results: Baseline approach.

2.2 Timbre feature extraction

Given the satisfying results of related work using Mel-frequency coefficients for high quality audio [2], we limit the timbre-based feature extraction to the average *MFCCs* 1-13. Instead of frame-wise extraction, we calculate the average for the *MFCC* values over each song in order to obtain compatibility with the global descriptors described below. For monophonic material, silent frames are estimated from the energy envelope and excluded from the feature extraction process. For polyphonic material, voiced sections are estimated from the pitch salience function obtained during the predominant fundamental frequency (**f0**) estimation [10].

2.3 Vibrato feature extraction

Vocal vibrato is defined as an oscillation of the (**f0**) within a rate of 4 to 8 Hz [15] and a depth of less than 200 cents. It seems convenient to estimate vibrato directly from the fundamental frequency curve as opposed to the application of time-domain filter banks as in Nwe et al. [6]. We obtain the pitch contour of the recordings using the *MELODIA* vamp plugin[1], which implements a state-of-the-art predominant fundamental frequency estimation algorithm [10] for monophonic and polyphonic audio material. After removing silences, the estimated f0 curve can be treated as a time series and high-pass filtered at 2Hz in order to obtain a zero-centered signal $\mathbf{x_0(t)}$ which preserves only fast pitch fluctuations. If vibrato is present, the spectrum of this signal $\mathbf{X_0(f)}$ contains a peak within the considered frequency range. The instantaneous vibrato rate corresponds to the peak frequency **f** and the vibrato depth **d** can be estimated as the amplitude differ-

[1] http://mtg.upf.edu/technologies/melodia

ence of the zero-centered pitch fluctuation in the current frame **[t₁; t₂]**:

$$d = \max(x(t))\max_{t1}(x(t)) - \min(x(t)) \quad (1)$$

In order to obtain global descriptors, we determine the average and the standard deviation of vibrato rate and depth over each track. Furthermore, the vibrato amount is defined as the relative number of frames per song containing vibrato.

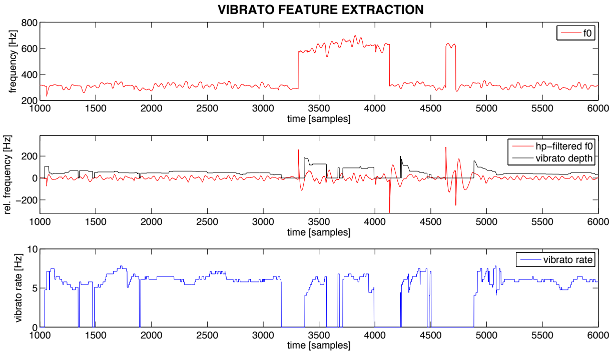

Figure 2. Extraction of vibrato features. *Top:* f_0 contour; *middle:* high-pass filtered f_0 contour and estimated vibrato depth; *bottom:* estimated vibrato frequency.

2.4 Performance descriptor extraction

In order to model a singer's characteristic performance style, we extract statistical features describing the behavior of pitch, dynamics and note duration. We use the transcription system described in Gómez and Bonada [11], which estimates a symbolic note representation from the $\mathbf{f_0}$ trajectories mentioned in the previous section. The analyzed recording is segmented into notes, which are represented with their respective duration, energy and pitch. Frequency values are quantized to an equal-tempered scale relative to the locally estimated tuning frequency. From the note representation we extracted the lowest and highest pitch, the pitch range as well as the standard deviation of the pitch value, the pitch fluctuation for each track. After normalizing the dynamics for each song, we furthermore extracted the average volume and average note duration and their respective standard deviations, the dynamic and duration fluctuation. The onset rate is determined as the number of notes per second. Comparing quantized $\mathbf{q_i}$ and non-quantized $\mathbf{u_i}$ pitch values, we define the maximum detuning $\mathbf{d_{max}}$ for each song:

$$d_{\max} = \max_i |q_i - u_i| \quad (2)$$

And its standard deviation, the tuning fluctuation. By calculating the absolute difference between the $\mathbf{f_0}$ envelope and the non-quantized note values $\mathbf{u_i}$ relative to the total number of frames **N** in each song, we estimate the ***amount of ornamentation*** **o** as follows:

$$o = \frac{\sum_{i=1}^{N} |f_{0i} - u_i|}{N} \quad (3)$$

A key characteristic of flamenco singing is a singer's ability to stretch phrases with melisma and ornamentation without breaking the note flow by breathing or pausing. Silences between phrases are chosen consciously and serve as an expressive factor. Consequently, the ***amount of silence*** estimated from the relative number of unvoiced frames per song and the maximum phrase length have been selected as corresponding descriptors.

Feature set	Descriptor
Timbre	Average MFCC 1-13
Vibrato	Average vibrato rate
	Vibrato standard deviation
	Average vibrato extend
	Vibrato amount
Performance	Lowest pitch
	Highest pitch
	Pitch range
	Pitch standard deviation
	Average volume
	Volume standard deviation
	Average note duration
	Note duration standard deviation
	Onset rate
	Maximum detuning
	Tuning fluctuation
	Amount of ornamentation
	Amount of silence

Table 2. Summary of extracted descriptors.

2.5 Learning task

Automatic singer identification refers to the task of automatically identifying a performer by analyzing descriptors extracted from audio recordings. Here, we classify among a set of possible candidates based the described features extracted on a song-level. Hence, we are interested in obtaining a classifier F of the following form

$$F(MusicFragment) \rightarrow Performer \quad (4)$$

where ***MusicFragment*** is the set of music fragments and ***Performer*** is the set of possible singers to be identified. Each music fragment is characterized by all of the different subsets of the extracted descriptors described above. After comparing various classifiers, we chose *Support Vector Machines (SVM)* [12] for the task due to the throughout consistent performance. Parameters are empirically adjusted to: complexity **c=15**, tolerance **t=0.001** and $\boldsymbol{\varepsilon = 10^{15}}$, using feature set normalization and a linear kernel.

2.6 Classification

We perform the learning task and classification task in the WEKA machine learning environment [13] applying a ten-fold cross-validation. For both monophonic and polyphonic material, we classify using the full feature set as well as single feature groups in order to compare their suitability for this task. An additional feature set is obtained by applying a SVM subset evaluation. We evaluate the algorithm performance by calculating the percentage of correctly classified instances (CCI), averaged over N=10 folds:

$$CCI = \sum_{i=1}^{N} \frac{\#\,\text{correctly classified instances}}{\#\,\text{instances}} \qquad (5)$$

3. EVALUATION AND RESULTS

3.1 Data

In order to evaluate and compare the two approaches described subsequently, we gather a total of three databases including monophonic and polyphonic flamenco collections and for comparative analysis a polyphonic Opera singing dataset. The database Fl-Mono contains a total of 65 recordings of flamenco a cappella songs by five renowned male artists. The collection contains recent as well as historical recordings with an overall strongly varying audio quality. The total amount of audio material sums to approx. 2 hours and 51 minutes. Covering a great variety of styles and eras, the polyphonic flamenco dataset Fl-Poly contains a total of 150 commercially available recordings by three male and two female renowned singers. The audio quality is throughout high and the instrumentation is limited to singing voice, guitar and percussive elements such as hand-clapping. For comparative reasons we furthermore gathered an opera singing collection Fl-Poly containing 150 tracks with full orchestra accompaniment from commercially available CDs. The dataset covers three male and two female professional singers with orchestral instrumentation.

3.2 Experiments and results

3.2.1 System performance

Feature set	Correctly classified instances (CCI)		
	Fl-Mono	*Fl-Poly*	*Op-Poly*
Timbre	60.0%	86.7%	70.5%
Vibrato	72.3%	63.7%	61.1%
Note	63.1%	53.3%	57.0%
All	80.0%	88.0%	66.4%
SVM subset	83.1%	86%	76.5%

Table 3. Classification results: Selected approach.

The evaluation results for the proposed algorithm tested on all three databases is illustrated in table 3. For the monophonic flamenco dataset, the singer identification based on averaged MFCCs obtains a slightly inferior result compared the baseline approach (65.00%) where MFCCs are extracted frame-wise. Vibrato and statistical note descriptors clearly outperform the timbre-based method. An increase in accuracy is observed for the combined feature set. Applying a SVM-based subset selection, in which the five lowest ranked features are discarded, results further improve. In contrast, for the high quality polyphonic audio recordings the timbre feature set gives a higher precision than vibrato and statistical note features. Obviously, the absence of strong spectral distortions and background noise allows a more precise timbre modeling. However, given the small size and lack of variety of both databases, these results might be influenced by the album effect described in the following section. Furthermore, polyphonic flamenco styles as well as score-based opera singing leave less interpretational freedom than spontaneous a cappella flamenco performances. Thus, statistical note descriptors are less informative regarding the performance style of a particular artist. A further bias may be introduced from the fact that the monophonic database contains male singers only, providing a more homogeneous voice timbre and consequently a more complex task of determining the singer. Also, using predominant f0 estimation to transcribe the singing voice from polyphonic audio is more susceptible to errors than for monophonic signals. Specially in Opera singing instruments might take over main melodic lines during interludes and cause errors in the extracted features. Nevertheless, compared to the timbre-based approach, a fusion of the feature sets improves the accuracy for polyphonic material. We observe a further increase in precision for monophonic flamenco and opera singing when applying an SVM subset selection.

3.2.2 Influence of the album effect

For small databases with a lack of variety regarding albums, the so-called album effect might occur: The homogeneous sound of the album determined by the production, mixing and mastering processes is modeled instead of the characteristics of the singing voice timbre. Consequently, tracks which do not originate from the albums contained in the training database are specially prone to misclassification. [14] observe this effect for an automatic singer identification based on timbre and vibrato descriptors as an elevated error rate for the case when tracks in training and test datasets originate from separate sources.

In the scope of this study, we explore the robustness towards the album effect by comparing two train-test setups for a simplified two-class problem. The album split represents a worst-case scenario where the model is trained on one album per singer while the test set originate from other sources. In the random split, the same tracks were randomly placed in training or test set, while keeping the sizes of the sets constant for both scenarios. The results clearly show a decrease in accuracy for the timbre-based classification when the model is trained on a single album per singer. The same effect occurs for the performance descriptors, for which the explanation is less obvious. Analyzing the information gain of the note descriptors regarding the classes, we noticed that the attributes ornamentation amount and onset rate obtained a high ranking for the train but not the test set in the album split. This is an indication for a lack of variety regarding styles within the data: Similar to the error of modeling an album specific timbre, in this case style specific characteristics might have been included in the model instead of singer specific features. When the full attribute set is used, the results for both scenarios are within the same range. This experiment clearly demonstrates the advantage of an incorporation of vibrato-based features for small training sets with little diversity regarding sources and styles.

Feature set	**Correctly classified instances (CCI)**	
	Album split	*Random split*
Timbre	41.7%	81.0%
Vibrato	87.5%	76.1%
Note	41.7%	95.2%
All	91.7%	90.5%

Table 4. Influence of the album effect.

4. CONCLUSION

Our experiments show that an identification based on spectral descriptors gives reliable results for databases containing high-quality audio recordings, even if the amount of available material is limited. On the other hand, this approach is not robust towards spectral distortions and varying overall timbre due to inconsistent recoding situations and audio quality. We therefore extend the feature set with vibrato descriptors extracted from the f0 envelope. In order to model the performance style of a particular artist, we extract global attributes describing temporal, melodic and dynamic behavior from automatic transcriptions. The resulting method combining vibrato, note and timbre descriptors shows an increase in accuracy for monophonic styles and robustness towards spectral distortions. We confirm the album effect for timbre-based approaches and furthermore observe a similar phenomenon for note descriptors when the database lacks a variety of styles.

Acknowledgments

This research has been partly funded by the Agencia de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) / Generalitat de Catalunya / 2013-AAD-01874 and the Spanish Ministry of Economy and Competitiveness (SIGMUS project, Subprograma de Proyectos de Investigacin Fundamental no Orientada, TIN2012-36650).

5. REFERENCES

[1] W. Cai, Q. Li and X. Guan, "Automatic singer identification based on auditory features", in *Proc. of the International Conference on Natural Computation*, Shanghai, 2011, pp. 1624-1628.

[2] W. Tsai and H. Lee, "Automatic Singer Identification Based on Speech-Derived Models," in *Proc. of the Int. conference on advancements in information technology*, 2012, pp. 13–15.

[3] M. Lagrange, A. Ozerov and E. Vincent, "Robust Singer Identification in Polyphonic Music Using Melody Enhancement and Uncertainty-Based Learning," in *Proc. of the 13th Int. Society for Music Information Retrieval Conference (ISMIR)*, Porto, 2012, pp.595-600.

[4] J. Shen, J. Sheperd, B. Cui and K.-L. Tan, "A novel framework for efficient automated singer identification in large music databases,", *ACM Transactions on Information System*, vol. 27, no. 3, pp. 1-13, 2009.

[5] H. Fujihara and M.Goto, "A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre similarity-based music information retrieval," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 638–648, 2010.

[6] T. L. Nwe and H. Li, "On Fusion of Timbre-Motivated Features for Singing Voice Detection and Singer Identification," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, 2008, pp. 2225–2228.

[7] R. Sridhar and T. V. Geetha, "Music Information Retrieval of Carnatic Songs Based on Carnatic Music Singer Identification," in *Proc. of International Conference on Computer and Electrical Engineering*, Phuket, 2008, pp. 407–411, IEEE.

[8] C. L. Hsu and J. S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 2, pp. 310–319, 2010.

[9] N. Kroher, *The Flamenco Cante: Automatic Characterization of Flamenco Singing by Analyzing Audio Recordings*, Master thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2013.

[10] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[11] E. Gómez and J. Bonada, "Towards Computer-Assisted Flamenco Transcription: An Experimental

Comparison of Automatic Transcription Algorithms as Applied to A Cappella Singing," Computer Music Journal, vol. 37, no. 2, pp. 73–90, June 2013.

[12] N. Christiani and J. Shawe-Taylor, "An introductionto support vector machines and other kernel-based learning methods," *Cambridge University Press*, 2000.

[13] M. Hall, H. National, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, no. 1, pp.10–18, 2009.

[14] T. L. Nwe and H. Li, "Exploring vibrato-motivatedacoustic features for singer identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 519–530, 2007.

[15] J. Sundberg, "Measurement of the vibrato rate of ten singers,“ *Quarterly Progress and Status Report, STL-QPSR,* vol.33, no. 4, pp.73-86, 1992.