Improving accompanied Flamenco singing voice transcription by combining vocal detection and predominant melody extraction

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

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

ABSTRACT

While recent approaches to automatic voice melody transcription of accompanied flamenco singing give promising results regarding pitch accuracy, mistakenly transcribed guitar sections represent a major limitation for the obtained overall precision. With the aim of reducing the amount of false positives in the voicing detection, we propose a fundamental frequency contour estimation method which extends the pitch-salience based predominant melody extraction [3] with a vocal detection classifier based on timbre and pitch contour characteristics. Pitch contour segments estimated by the predominant melody extraction algorithm containing a high percentage of frames classified as non-vocal are rejected. After estimating the tuning frequency, the remaining pitch contour is segmented into single note events in an iterative approach. The resulting symbolic representations are evaluated with respect to manually corrected transcriptions on a frame-by-frame level. For two small flamenco dataset covering a variety of singers and audio quality, we observe a significant reduction of the voicing false alarm rate and an improved voicing F-Measure as well as an increased overall transcription accuracy. We furthermore demonstrate the advantage of vocal detection model trained on genre-specific material. The presented case study is limited to the transcription of Flamenco singing, but the general framework can be extended to other styles with genre-specific instrumentation.

1. INTRODUCTION

Flamenco, a music tradition with origins in Andalusia, Southern Spain, has mainly germinated and nourished from the singing tradition [1] and until now the voice remains its central element, usually accompanied by the guitar and percussive elements. As an oral tradition, performances are typically spontaneous and not score-based. Melodic skeletons and lyrics are passed orally from generation to generation and only rare manual annotations exist. With flamenco gaining in popularity around the world and the growing interest of musicians with a formal music education in the genre, the transcription of a large number of performances into a symbolic note representation has become of importance, not only for musicological studies but also for educational purposes. Furthermore, accurate note transcriptions can be used to compute descriptors modeling a specific performance or performance style. Such features are used in related music information retrieval (MIR) tasks, such as automatic singer identification or melodic similarity [11]. Obtaining an automatic transcription of the singing voice melody directly from an audio file comprises two steps: First, the estimation of the fundamental frequency of the singing voice and second, the segmentation into single note events. For the task of transcribing accompanied Flamenco singing, a recent approach [2] based on predominant melody extraction gives satisfying results when comparing to manually corrected reference scores. However, the authors report false positives in the voicing detection as a main source of error: The polyphonic predominant melody extraction algorithm [3] estimates voiced sections mainly based on pitch salience and continuity and consequently transcribes the instantaneous perceptually dominant melody. During the intro and short instrumental interludes, the guitar carries the main melody and is in some cases mistakenly transcribed. In order to reduce this type of error, it is necessary to automatically distinguish between guitar and vocal segments. The presented approach combines predominant melody extraction with a frame-wise vocal/non-vocal classification, rejecting contour segments with a high percentage of predicted non-vocal frames. The classifier is adapted to the case of Flamenco singing accompanied by guitar. Nevertheless, for other styles, specially with genre-specific instrumentation, the same framework can be applied by training the classifier on the given material.

2. SCIENTIFIC BACKGROUND

The automatic generation of a symbolic note representation of the singing voice melody from a piece of audio represents a challenging task in MIR, gaining in complexity for polyphonic music signals and expressive singing styles. The flamenco singing voice can be described as highly unstable in pitch, tuning and timbre and performances are characterized by a large amount of spontaneous melisma and ornamentations. These features ag-

gravate the process of automatic transcription, specially regarding note segmentation, as demonstrated in [4], where significantly higher accuracies are achieved for transcribing jazz vocal pieces compared to flamenco recordings. Nevertheless, previous approaches give promising results with overall accuracies within a tolerance of 50 cents of 70% for monophonic [4] recordings estimating the fundamental frequency from the spectrum autocorrelation and 85% for polyphonic [2] recordings using a predominant melody extraction algorithm [3]. As mentioned above, for the case of accompanied Flamenco singing, errors mainly originate from mistakenly transcribed guitar sections. We therefore incorporate a model-based vocal detection, which classifies on a frame-level based on timbre and pitch contour characteristics. Vocal segment detection algorithms [5-10] using machine learning models have previously shown to give convincing results with frame-wise accuracies of up to 87% [10] and to improve the performance of singing voice related MIR tasks, such as automatic singer identification [5]. In a comparison of acoustic features for this task [9], the Mel-frequency coefficients have given the highest accuracy. Other approaches include descriptors related to vibrato [5], harmonic content [5,6] and pitch contour characteristics [5, 6, 8, 10].

3. TRANSCRIPTION METHOD

The basic framework of the transcription algorithm corresponds to the one described in [2], extended by the model-based vocal detection system.

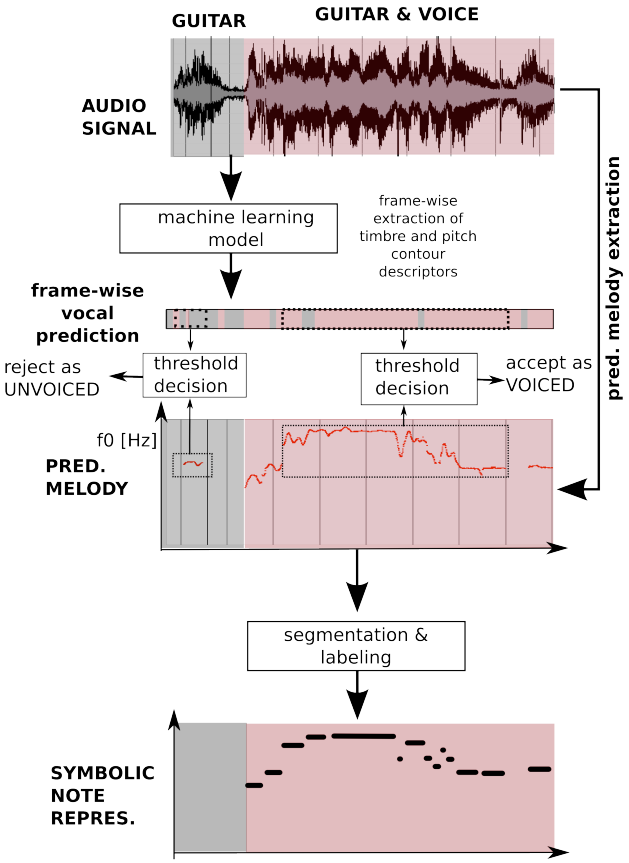

Figure 1. Singing voice transcription framework.

The block diagram in figure 1 gives an overview of the process, the various stages are described below in the further detail.

3.1 Predominant melody extraction

We apply a state-of-the-art predominant melody extraction as described in detail in [3]: After applying a perceptually-based filter, peaks are extracted from the short-term spectrum and used to compute a salience value based on harmonic summation for all possible fundamental frequency values within an adjustable frequency range [$\mathbf{f_{min}}$, $\mathbf{f_{max}}$]. The final $\mathbf{f_0}$-curve is estimated by contour tracking based on auditory streaming principles. With the aim of rejecting contours not belonging to the melody, the algorithm evaluates the *contour mean salience* $C_{\bar{s}}$

$$C_{\bar{s}} = \frac{\sum s(f(t))}{T} \quad (1)$$

for a salience value **s** and a contour **f(t)** of length **T** against a threshold **τ**, calculated from mean and standard deviation of the contour mean salience $C_{\bar{s}}$ as:

$$\tau = \bar{C}_{\bar{s}} - v * \sigma(C_{\bar{s}}) \quad (2)$$

The corresponding adjustable parameter, which controls the voicing threshold is referred to as *voicing tolerance* **v**. As shown in [2], the overall accuracy of the voicing detection can be significantly improved by optimizing both considered frequency range and voicing tolerance for a given song. In order to avoid manual adjustment and further improve the voicing detection accuracy, we set the parameters to empirically determined values ($\mathbf{f_{min}}$=120Hz, $\mathbf{f_{max}}$=720Hz, **v**=0.2) and combine the voicing detection process with a model-based approach described below.

3.2 Vocal detection

Taking advantage of prior knowledge about the limited genre-specific instrumentation, we aim to train a classifier based on timbre, harmonic spectrum and pitch contour characteristics in order to automatically detect vocal segments.

3.2.1 Feature extraction

Based on the good performance reported in [9], we extract the *Mel-frequency cepstral coefficients* (*MFCCs*) *1-13* to model overall timbre. We furthermore calculate the total pitch salience s_{Total} of all M estimated spectral peaks frequencies in the current frame t:

$$s_{total}(t) = \sum_{m=1}^{M} s(f_m(t)) \quad (3)$$

Based on the observation that melodic contours of vocal segments, specially in the case of Flamenco singing, are characterized by fast pitch fluctuation (figure 2) originating from vocal vibrato and ornamentations, we furthermore extract the pitch standard deviation $\mathbf{p_{StDev}}$ of the current estimated predominant melody contour **f(t)**:

$$p_{StDev}(t) = \sigma(f(t)) \tag{4}$$

3.2.2 Attribute selection

In order to determine the most suitable features for this task among the descriptors described above, we perform an attribute ranking based on information gain in the *WEKA* machine learning environment and chose the six highest ranked features: $\mathbf{p_{StDev}}$, $\mathbf{s_{Total}}$, **MFCC1**, **MFCC3**, **MFCC5** and **MFCC7**.

3.2.3 Classifier

We train a linear support vector classifier as described in [13] using the *liblinear* [14] library. We empirically adjust the cost parameter to **c**=1.0 and tolerance of the termination criterion to **ε**=0.01.

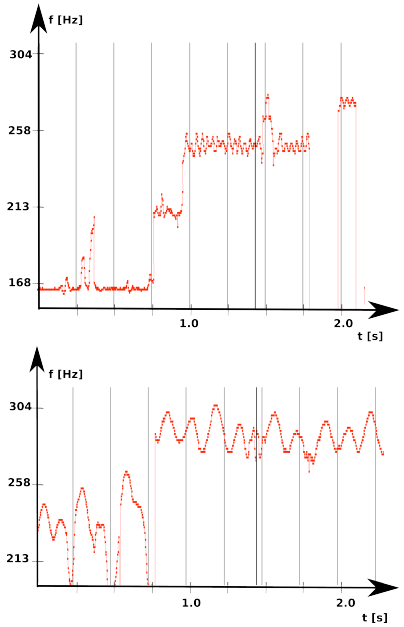

Figure 3. Pitch contour of a guitar (top) and a voice .

3.2.4 Final contour selection

We define an adjustable threshold **th** as the percentage of frames in a contour segment classified as unvoiced above which the contour is discarded for further processing. Consequently, a contour of length **M** with a frame prediction **y(m)**=0 for unvoiced and **y(m)**=1 for voiced frames is rejected, if:

$$\frac{\sum_{m=1}^{M} y(m)}{M} * 100 > th \tag{5}$$

As shown in section 4, we empirically optimize this threshold for the given dataset to **th**=80%.

3.3 Note segmentation and labeling

Below we provide a summary of the basic steps implemented in our transcription system to segment the estimated fundamental frequency envelope into single note events. For a more detailed description, we refer to [4].

3.3.1 Tuning estimation

The tuning is assumed to be constant throughout the analyzed excerpt. This assumption is considered to be valid for accompanied Flamenco singing, since in contrary to a cappella styles, where tuning may vary strongly, singers tend to adjust to the tuning of the guitar. Nevertheless, the absolute tuning frequency is not known. We therefore compute a histogram of instantaneous pitch deviations from the equal tempered scale with a reference tuning of 440 Hz. The maximum of the histogram corresponds to the shift in tuning with respect to the reference and the tuning frequency for the analyzed excerpt is computed accordingly.

3.3.2 Note segmentation

After adjusting the tuning frequency, the fundamental frequency contour is segmented into short notes in a dynamic programming approach by maximizing the likelihood among all possible note progressions. We define a possible note progression path **p** as

$$p = [n_0, n_1, ... n_{N-1}] \tag{6}$$

where each note $\mathbf{n_i}$ contained in the path is characterized by its start time, pitch and duration. Restrictions in pitch result from the pitch range estimated from the fundamental frequency contour. Possible paths are furthermore limited by a priori defined minimal and maximal note durations as well as the excludability of temporal overlap. For each note and each note transition, a likelihood is computed based on duration, pitch with respect to the instantaneous pitch estimate, low-level feature analysis and the existence of voiced frames within a segment. The overall likelihood **L(p)** of a possible path of length **N** is calculated as the product of note $\mathbf{L(n_i)}$ and note transition likelihoods $\mathbf{L(n_{i-1},n_i)}$ contained in the path:

$$L(p) = L(n_0) * \prod_{i=1}^{N-1} L(n_i) * L(n_{i-1}, n_i) \tag{7}$$

3.3.3 Short note consolidation and tuning refinement

Since notes may be longer than the maximum duration assumed for the likelihood estimation, sequences of short notes with the same pitch are consolidated if the dynamics of low-level features do not indicate a long note instead of a series of note onsets. In an iterative approach we repeat this process after re-estimating the tuning fre-

quency. In this step, the tuning deviation histogram is computed from the estimated instantaneous fundamental frequency weighted by the assumed note duration.

4. EVALUATION

We first evaluate the frame-wise accuracy of the vocal detection model for three small datasets in a ten-fold cross-validation. Subsequently, we compare the frame-wise voicing detection accuracy of the predominant frequency estimation with and without incorporation of the model-based classification with respect to manually annotated data. In order to investigate the advantage of training on genre-specific material, we compare the voicing detection performance for a vocal model trained on flamenco songs with a model trained on Western commercial music. Since the transcription process takes the previous and subsequent note values into account, we furthermore compare the overall accuracy of the resulting note representations based on manually corrected transcriptions to determine in how far a reduction of false positives influences the overall transcription performance.

4.1 Databases

The voicing detection is tested on two datasets: *FL_FULL* and *FA_EXC*. *FL_FULL* contains ten full songs with an overall duration of approximately 40 minutes and covers a variety of flamenco subgenres *FA_EXC* contains 40 excerpts of accompanied Flamenco singing of approximately one minute each. The excerpts cover mainly the sections where the singing voice is present and do not include guitar intros or longer interludes. All songs belong to the same style (*Fandagos*). Both databases cover a variety of renowned male and female singers and guitarists. The recording quality strongly varies among the excerpts. In order to investigate the advantage of training on genre-specific material, we train a vocal detection model on a third dataset *WEST_FULL*, containing ten western commercial music recordings with vocals and varying instrumentation. All audio files are sampled at 44.1 kHz with a resolution of 16 Bit. As a ground truth for evaluation purposes, the audio files contained in all three databases were manually segmented into voiced and unvoiced frames. For the database *FA_EXC* we furthermore had access to manually corrected transcriptions based on the automatic transcriptions obtained with the algorithm described above. The correction was conducted by a trained musician without special knowledge of Flamenco, following general guidelines regarding transcription of ornamentation details and pitch glides set by Flamenco experts.

4.2 Evaluation methodology

We evaluate the frame-wise accuracy of the voicing detection by calculating the *voicing recall* (% of all voiced frames correctly estimated as voiced), *voicing false alarm* (% of all unvoiced frames mistakenly estimated as voiced), and *voicing precision* (% all frames classified as voiced which actually are voiced). We furthermore compute the F-measure as follows**:**

$$F = \frac{2 * precision * recall}{precision + recall} \tag{8}$$

In order to avoid over-fitting, all experiments described below are conducted in a song-wise ten-fold cross-validation. For each of the measures mentioned above, we give the average value among all folds.

4.3 Results

4.3.1 Accuracy of the model-based vocal detection

In a first experiment we evaluate the performance of the vocal detection model by means of *correctly classified instances (CCI)*. For all three databases, the evaluation was conducted in a frame-based ten-fold cross-validation process

4.3.2 Frame-wise voicing detection accuracy

We analyze the voicing detection performance with and without the model based vocal detection (VD) with respect to manually corrected vocal/non-vocal segmentation. Table 1 gives the average performance measures among the song-wise ten-fold cross-validation. The decision threshold for the rejection of a contour was set to **th**=80%.

	FA-EXC		*FL_FULL*	
	Without VD	With VD	Without VD	With VD
Precision	84%	90%	71%	83%
Recall	97%	95%	95%	92%
False alarm	185	8%	30%	15%
F-measure	0.89	0.92	0.81	0.87

Table 1. Voicing detection accuracy with and without model-based vocal detection.

For the database *FL_FULL* the results show a significant reduction of false positives of 15% as well as an improvement in precision by 12%. The small decrease in the voicing recall 3% indicates that a number of voiced frames has been mistakenly rejected during the vocal detection stage. Nevertheless, the overall F-measure improves from 0.81 to 0.87. We observe the same trend for the database *FA_EXC*, but with a lower increase in performance. This can be explained by the fact that the excerpts mainly cover singing voice sections and therefore the false alarm rate without the incorporation of a vocal detection is comparatively low.

4.3.3 Influence of the decision threshold

In order to adjust the threshold for the rejection of a given contour, we computed the F-measure for a single fold of the database *FA_EXC* in dependency of the parameter **th**.

	F-Measure
No vocal detection	0.860
th=50%	0.900
th=60%	0.917
th=70%	0.925
th=80%	0.929
th=90%	0.927
th=99%	0.925

Table 2. Voicing detection accuracy for varying decision thresholds.

We accordingly adjust the decision threshold to **th**=80%.

4.3.4 Influence of the training database

As many other genres, flamenco music is characterized by a typical instrumentation, namely vocals, guitar and hand-clapping. In order to investigate the advantage of training the vocal detection model on such genre-specific data compared to a generalized vocal detection model, we additionally train a model on the *WEST_FULL*. For each fold of the database *FL_FULL* we compare the performance for incorporating the genre-specific model trained on the current fold vs. the general model trained on *WEST_FULL*.

	Genre-specific VD model	General VD model
Voicing false alarm	15%	8%
Voicing precision	83%	81%
Voicing recall	92%	59%
F-measure	0.87	0.57

Table 3. Voicing detection accuracy for different training databases.

The results given in table 3 clearly show a significant decrease in performance when classifying based on the generalized model. The F-measure even drops below the value of 0.81 without vocal detection. It becomes obvious, that at least for such small amount of training data, the vocal detection model needs to be trained on genre-specific data.

4.3.5 Accuracy of the note transcription

For the database *FA_EXC*, we furthermore evaluate the resulting note transcriptions by means of *overall accuracy* (% of correctly estimated frames, including voicing detection) and *raw pitch accuracy* (% of voiced frames for which the estimated pitch is within a range of 50 cents). Again, the results refer to the average measures over a 10-fold cross-validation.

The results show an improvement in the overall accuracy of approximately 2%. The raw pitch accuracy is slightly lower when the vocal detection model is incorporated in the system. This probably results from a small number of falsely rejected voiced frames. Consequently, the reduction of false positives does not seem to have a strong influence on the note segmentation and labeling process for the case of the dataset containing mainly voiced frames. A larger improvement can be expected for a dataset containing longer unvoiced sections and guitar interludes. We therefore plan a ground truth transcription for the full dataset *FL-FULL* for future work.

	Without VD model	With VD model
Raw pitch accuracy	67.81%	67.06%
Overall accuracy	88.29%	90.37%

Table 4. Transcription accuracy with and without model-based vocal detection.

5. CONSLUSIONS

We propose the incorporation of a statistical learning model into the automatic singing voice transcription system described in [2] in order to reduce the number of voicing false positives. We train a classifier on a small number of frame-wise extracted timbre, harmonic and pitch contour descriptors and assume a genre-specific instrumentation. A contour segment estimated by the predominant melody extraction algorithm is rejected if more than 80% of the contour frames are classified as unvoiced. The results clearly indicate that incorporating the model-based vocal detection significantly reduces the percentage of voicing false positives and improves the voicing detection F-measure. By comparing the performance for classifiers trained on genre-specific and Western commercial music databases, we confirm the advantage of the former, at least for a small number of training instances. Furthermore, we observe a small increase in the overall transcription accuracy resulting from the reduction of mistakenly transcribed guitar notes. A larger improvement can be expected on a dataset containing larger unvoiced section, which will be included in subsequent work.

Acknowledgments

This research has been partly funded by the Spanish Ministry of Economy and Competitiveness (SIGMUS project, Subprograma de Proyectos de Investigacin Fundamental no Orientada, TIN2012-36650) and the PhD fellowship program of the Department of Information and Communication Technologies of Universitat Pompeu Fabra, Barcelona, Spain.

6. REFERENCES

[1] A. Álvarez Caballero. *El cante flamenco*. Alianza Editorial, Madrid, 2004.

[2] J. Salamon, J. Bonada, P. Vera and P. Cabañas, "Predominant Fundamental Frequency Estimation vs. Singing Voice Separation for the Automatic Transcription of Accompanied Flamenco Singing," in *Proc. of the 3rd Interdisciplinary Conference on*

Flamenco Research and the 2nd International Workshop on Folk Music Analysis. Seville, 2012.

[3] 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, 1759–1770, 2012.

[4] 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 Jounal*, vol. 37, no. 2, pp. 73-90, 2013.

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

[6] V. Rao, C. Gupta and P. Rao, "Context-aware features for singing voice detection in polyphonic music," In *Proceedings of the 9th International Workshop on Adaptive Multimedia Retrieval*. Barcelona, Spain, 2011.

[7] I. Leonidas, J.-L. Rouas, "Exploiting Semantic Content for Singing Voice Detection," In *Proc. of IEEE Sixth International Conference on Semantic Computing*, Palermo, Italy, 2012, pp.134–137.

[8] M. Rocamora and A. Pardo, "Separation and Classification of Harmonic Sounds for Singing Voice Detection," In *Proc. of 17th Iberoamerican Congress, CIARP*, Buenos Aires, Argentina, 2012, pp. 707–714.

[9] M. Rocamora, and P. Herrera, "Comparing audio descriptors for singing voice detection in music audio files," In *Proc. of Brazilian Symposium on Computer Music*, 2012.

[10] M. Mauch, H. Fujihara, K. Yoshii and M. Goto, "Features for the Recognition of Vocal Activity and Instrumental Solos in Polyphonic Music," In *Proc. of the International Society of Music Information Retrieval Conference (ISMIR)*, Miami, Florida, 2011, pp. 233-238.

[11] N. Kroher. *Automatic characterization of Flamenco Singing by Analyzing Audio Recordings*. Master thesis, Universitat Pompeu Fabra, Barcelona, Spain.

[12] 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, 2009, pp. 10-18.

[13] K. Crammer and Y. Singer, "On the learnability and design of output codes for multi-class problems.," In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, 2000.

[14] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of machine learning research*, vol. 9, 2008, pp. 1871-1874.