Automatic Transcription of Flamenco Singing From Polyphonic Music Recordings

Nadine Kroher and Emilia Gómez

***Abstract*—Automatic note-level transcription is considered one of the most challenging tasks in music information retrieval. The specific case of flamenco singing transcription poses a particular challenge due to its complex melodic progressions, intonation inaccuracies, the use of a high degree of ornamentation, and the presence of guitar accompaniment. In this study, we explore the limitations of existing state of the art transcription systems for the case of flamenco singing and propose a specific solution for this genre: We first extract the predominant melody and apply a novel contour filtering process to eliminate segments of the pitch contour which originate from the guitar accompaniment. We formulate a set of onset detection functions based on volume and pitch characteristics to segment the resulting vocal pitch contour into discrete note events. A quantised pitch label is assigned to each note event by combining global pitch class probabilities with local pitch contour statistics. The proposed system outperforms state of the art singing transcription systems with respect to voicing accuracy, onset detection, and overall performance when evaluated on flamenco singing datasets.**

***Index Terms*—Automatic music transcription, Music information retrieval, Singing voice, Pitch contour, Audio Content Description.**

I. Introduction

A. Definition and Motivation

FLAMENCO music is a rich improvisational art form with roots in Andalusia, a province in southern Spain. Due to its particular characteristics and importance for the cultural identity of its area of origin, flamenco as an art form was inscribed in the UNESCO List of Intangible Cultural Heritage of Humanity[1] in 2010. Given the growing community of flamenco enthusiasts around the world, a need for computational methods to aid study and diffusion of the genre has been identified. First efforts have been made to adapt existing music information retrieval (MIR) techniques to the specific nature of flamenco [1]. Having evolved from an a cappella singing tradition [11], the vocals represent the central element of flamenco music, accompanied by the guitar, percussive hand-clapping and dance. Consequently, the main focus is set on developing algorithms which target the analysis of the singing voice.

Flamenco is a strongly improvisational and sparsely documented oral tradition, where songs and techniques have been passed from generation to generation. Given the resulting lack of scores, studies often rely on scant, labour-intensive manual transcriptions. In this study, we present a novel system for automatic singing transcription from polyphonic music recordings targeting the particular case of flamenco. The resulting automatic transcriptions are essential to a number of related MIR tasks, such as melodic similarity characterisation [3], similarity-based style recognition [2], singer identification [4] or melodic pattern retrieval [40] and can furthermore aid a broad variety of musicological studies [5].

Automatic singing transcription (AST) refers to the extraction of a note-level representation corresponding to the singing melody directly from audio recordings. This task comprises two main challenges: First, the estimation of the fundamental frequency corresponding to the sung melody (*vocal pitch extraction*) and second, its conversion into discrete note events (*note transcription*). In the resulting symbolic representation, each note is described by its onset time, duration and a pitch value, usually quantised to the equal tempered scale. While the generalised task of automatic music transcription (AMT) [6], [7] is considered a major challenge in MIR, the genre under study poses a number of additional difficulties for both, the vocal pitch extraction and the note transcription stage. Pitch estimation is a well-studied problem in MIR with a variety of sub-tasks such as multi-pitch and predominant melody extraction [8]. The difficulty in the context of flamenco singing is to extract the pitch contour corresponding to the vocal melody while omitting contour segments which originate from the guitar accompaniment. The voice is usually not present throughout the entire song but alternates with interludes in which the guitar takes over the main melodic line. Note transcription can be a trivial task when, depending on the instrumentation, note onsets coincide with significant pitch and volume discontinuities. The singing voice on the other hand, given its non-percussive and pitch-continuous nature, poses a particular challenge when segmenting a pitch contour into discrete note events. Flamenco singing is characterised by a large amount of melodic, partly micro-tonal, ornamentations, extensive use of vibrato and instability of timbre and dynamics. Furthermore, vocal melodies are often composed of a succession of conjunct degrees and singers tend to intonate significantly above or below target notes

Manuscript received August 16, 2015; revised December 13, 2015; accepted February 13, 2016. Date of publication February 18, 2016; date of current version March 23, 2016. This work was supported in part by the Ph.D. Fellowship of the Department of Information and Communication Technologies, Universitat Pompeu Fabra and in part by the projects SIGMUS (TIN2012-36650) and COFLA II (P12-TIC-1362). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hirokazu Kameoka.

N. Kroher was with the Music Technology Group, Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona 08018, Spain. She is now with the Department of Applied Mathematics II, University of Seville, Seville 41092, Spain (e-mail: nkroher@us.es).

E. Gómez is with Music Technology Group, Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona 08018, Spain (e-mail: emilia.gomez@upf.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2531284

[1]http://www.unesco.org/culture/ich/en/RL/flamenco-00363

[10]. Consequently, the complexity of onset detection and pitch labelling significantly increases and requires an algorithmic design which considers the aforementioned characteristics.

B. Related Work

Approaches to automatic note-level transcription from audio recordings date back as far as 1977 [12] and are extensively reviewed by [6] and [7]. Most systems described in these reviews provide a generic transcription framework covering a wide range of instruments and musical genres. As mentioned in [7], specific instruments exhibit particular characteristics which might not be captured by a generic transcription system.

Approaches dealing with the specific task of singing transcription have usually been developed in user-oriented MIR tasks such as query-by-humming (QBH), query-by-singing (QBS), sight-singing tutors or computer-aided composition tools. Such systems mainly assume user-input, i.e. the user singing a query, and are consequently designed to transcribe unaccompanied recordings sung by amateur singers. In this case, the vocal pitch extraction stage of the transcription algorithm is reduced to a monophonic pitch estimation problem and systems mainly differ in the note segmentation stage: [13] and [14] obtain an initial detection of note onsets directly from discontinuities in the volume envelope and [15] use a detection function based on changes in spectral band energies. There are obvious limitations to these rather simple segmentation approaches: Consecutive notes sung in legato may have a stable volume envelope and accompanying instruments may cause sudden spectral variation without a singing voice onset being present. A first approach entitled *island building* based on pitch contour characteristics was proposed in [16] and still finds application in more recent systems [13], [14]: Consecutive frames with a pitch estimate within a limited range are grouped into notes events. In a comparative study [17] a variety of pitch contour based segmentation approaches are compared, including adaptive filters, maximum likelihood estimation, probabilistic modelling and local quantisation. A recent approach to note segmentation [21], formulates interval transition as a hysteresis curve and locates note onsets based on the local cumulative pitch deviation.

Addressing the more complex task of singing transcription from polyphonic recordings [18], a multiple fundamental frequency estimator with an accent signal is proposed extract the dominant pitch trajectory. The segmentation stage relies on a probabilistic note event model, using a Hidden Markov Model (HMM) trained on manual transcriptions. In an extension to this approach [19], note transition probabilities were incorporated in the computational model. Recently, this note segmentation method was implemented in the computer-aided note transcription tool *tony* [20]. It should be mentioned that such probabilistic methods require a large amount of ground truth data during the training stage. This involves time-consuming manual annotations in particular for improvisational and strongly ornamented singing performances, as it is the case for flamenco music.

A recent trend in music information retrieval has focused on culture-specific non-Western music traditions [22] and has led to the adaptation of existing MIR approaches as well as the development of novel genre-specific techniques. It has been shown that underlying musicological assumptions, i.e. regarding rhythm or tonality, do not necessarily hold for non-Western music styles. In the context of flamenco singing, a first approach for monophonic transcription was proposed in [2]: Based on a contour simplification algorithm [27], an estimated monophonic pitch track is converted into a set of constant segments within which the absolute error between the pitch track and the fitted constant does not exceed a pre-determined threshold. A system for computer-assisted transcription of single-voiced a cappella singing recordings has been proposed in [23]: Given the absence of accompaniment, a monophonic pitch estimator based on spectral auto-correlation (SAC) represents the front-end of the system. The note segmentation stage is based on a likelihood maximisation method [24]: A dynamic programming (DP) algorithm is used to find the best among all possible notes segmentations along the entire track. The resulting short note transcription is refined in a post-processing stage, containing an iterative tuning estimation and short note consolidation process. The system was extended [25], [41] to the transcription of sung melodies from polyphonic flamenco recordings by replacing the monophonic pitch estimator with a predominant pitch extraction algorithm [26].

Nevertheless, the authors report mistakenly transcribed guitar notes as a main source of error and the inaccurate vocal detection as the major limitation of the system performance. The reported note transcription accuracies reported in [23] with a note f-measure of slightly below 0.4 furthermore indicate the difficulty of transcribing flamenco singing. Significantly higher performance is achieved when evaluating the same transcription algorithm on a dataset containing pop- and jazz singing excerpts. We therefore identify a need for improving the state of the art on flamenco singing transcription and design an algorithm robust towards the particular characteristics of the genre which is furthermore suitable for the analysis of accompanied flamenco singing recordings.

C. Contributions and Paper Outline

We propose a novel AST system capable of transcribing strongly ornamented improvisational flamenco performances and achieving higher accuracies than state of the art methods. Similar to Gómez *et al.* [25], we use a predominant melody extraction algorithm to extract the vocal pitch contour. We extend this method in two ways: In a pre-processing stage, we use spectral characteristics to select the stereo channel in which the vocals are more dominant for further processing. Subsequent to the predominant melody extraction, we apply a novel contour filtering method, in which pitch contour segments corresponding to guitar melodies are eliminated. In the note transcription stage, we formulate a set of novel onset detection functions based on pitch contour and volume characteristics, which prove robust towards the influence of the accompaniment as well as a high degree of melodic ornamentations. A pitch label is assigned to each note event by combining overall chroma probabilities with local pitch statistics. We refine the resulting symbolic transcription by applying musicological constraints regarding note duration and pitch range.

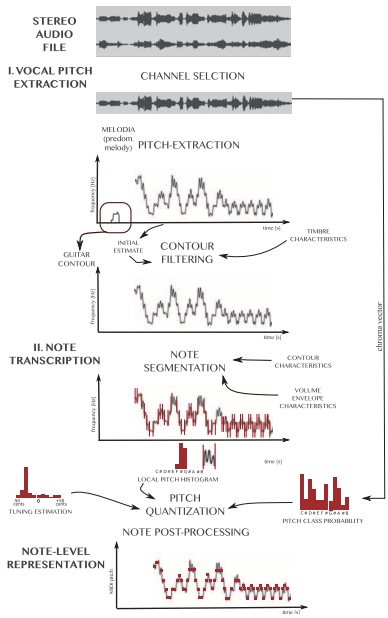

Fig. 1. Block diagram of the proposed system.

The remainder of the paper is organised as follows: We provide a detailed description of the proposed system in Section II and specify the evaluation methodology in Section III. We give experimental results in Section IV and finally conclude the study in Section V.

II. Proposed Method

The processing blocks of the proposed system are depicted in Figure 1. As mentioned above, automatic singing transcription systems consist of two main processing blocks: A vocal pitch extraction algorithm and a note transcription stage. Both stages of the proposed algorithm are described in detail below.

A. Vocal Pitch Extraction

In the context of singing voice transcription, we use the term *voicing* to describe whether the vocals are present in a frame or not. Consequently, for both monophonic and polyphonic transcription systems, the pitch values of *unvoiced* frames need to be eliminated in order to obtain an accurate time-frequency representation of the vocal pitch. For monophonic singing transcription systems, the task of vocal pitch extraction is reduced to a monophonic pitch estimation problem, since the singing voice represents the only instrument present. Nevertheless, silences and background sounds can cause noisy sections in the pitch track during unvoiced frames. While [13] and [14] locate unvoiced frames solely based on the volume envelope, in [21] a tree classifier based on a set of low-level descriptors in used. For polyphonic recordings, where the singing voice is accompanied by one or more instruments, the task of vocal pitch extraction gains in complexity due to the presence of various harmonic sources. In flamenco music, the singing voice is the dominant element accompanied by the guitar, which takes over the main melodic line during interludes. An obvious approach to extracting the vocal pitch would be to isolate the voice signal by applying source separation methods and using a monophonic pitch estimator on the resulting voice signal. A different strategy was applied in [25], where a predominant melody extraction algorithm is used to estimate the vocal melody. This method gives convincing results for sections of a track where the vocals represent the dominant music element. The same study has furthermore shown that the obtained overall performance regarding voicing and raw pitch accuracy is superior to a source separation approach. Nevertheless, mistakenly transcribed guitar contour segments during interludes, where the voice is absent and the guitar takes over the main melodic line, cause a relatively high number of voicing false positives.

Based on these prior findings, we adopt the predominant melody approach and extend it in order to reduce the number of mistakenly transcribed guitar pitch contour segments.

We first apply a *channel selection* in order to exploit the fact that in flamenco stereo recordings the vocals are usually more dominant in one of the two channels. We use only this channel for further processing. In pop music productions, it is common practice to place the vocals in the centre of the panorama during the mixing process. In flamenco recordings on the other hand, the vocals often appear stronger in one stereo channel. One reason for this phenomenon is that many such recordings are live stereo recordings, where the resulting panorama distribution corresponds to the physical location of singer and guitarist on stage. Even in multi-track flamenco productions it is a tradition to separate the voice and guitar in the artificially created panorama. An exception are productions with an extended instrumentation, i.e. two guitars or additional instruments, which is rather uncommon, or the obvious case of mono recordings. We want to exploit this fact by automatically selecting the channel with the stronger presence of the vocals for further processing in order to reduce the influence of the guitar during the vocal pitch extraction stage. We would like to point out, that while this channel selection improves system performance as shown in Section IV, our system does not rely on having a panorama separation of the sources. We simply exploit this fact whenever it is the case for a given track.

We then proceed to the *predominant melody extraction* from the selected channel. This process refers to the task of estimating the pitch track corresponding to the perceptually dominant melody in a given polyphonic music recording. It furthermore includes the task of determining if the main melody is present in a given time frame or not. In the scope of this paper we will refer to frames in which the main melody is estimated to be present as *melody frames* and all remaining frames as *non-melody frames*. Furthermore, we define a *contour* as a sequence of consecutive melody frames. We extract the predominant melody from the previously selected channel using the algorithm proposed in [26]. We selected this particular method due its available implementation in the *essentia* library [31] and in order to obtain a direct comparison to the study presented in [25], where it was also used as the front end of a flamenco singing transcription. We would like to mention that this method can be replaced by any other predominant melody extraction algorithm. While in the general task of pitch extraction, the performance is mainly limited by pitch estimation errors, i.e. octave errors,

the limitation in the scope of vocal pitch estimation is to a large extend conceptual: Even a perfect predominant melody extraction algorithm will estimate contours originating from the guitar, since the guitar accompaniment represents the perceptually dominant element during certain sections of a flamenco song.

After extracting the predominant melody, we apply a novel *contour filtering* process and eliminate parts of the pitch contour which are located outside the vocal sections. Vocal detection has been explored mainly as a machine learning tasks [28]–[30]. While good results have been obtained on a frame-level, such algorithms require a large amount of manually annotated training data. Here, we propose a novel vocal detection algorithms which works on a track level without any previous training stage and eliminates guitar contour segments based on spectral characteristics. The proposed scheme is based on two assumptions: First, the majority of all melody frames are voiced, or, in other words, the main melody largely corresponds to the vocal melody. This assumption is supported by the experiments carried out by Gómez *et al.* [25], where only around 15% of all estimated pitch values originate from the guitar. Second, we assume that guitar and vocal sections can be discriminated based on spectral characteristics. In preliminary experiments, we investigated a variety of spectral descriptors and found that the energy of the first 12 bark bands [32] are best-suited for this task (Figure 3 (a) and (b)). Based on these assumptions, we model melody frames originating from the guitar as outliers in a Gaussian distribution of bark band energies extracted for all melody frames.

Below, we provide a detailed description of the three stages involved in the vocal pitch extraction process and their implementation.

1) Channel Selection: The automatic channel selection is based on the fact that guitar and voice differ in their distribution of spectral energy (Figure 2): When the vocals are present, we observe an increase in the frequency range of 500 Hz to 6 kHz. We therefore select the stereo channel with a higher average presence in this range. For both audio channels with a sampling rate of $f_s = 44.1$ kHz, we compute the Short Time Fourier Transform (STFT) for a window size $N = 4096$ samples multiplied with a hanning window function, a zero padding factor of $m = 2$ and a hop size of $h_s = 1024$ samples. Accordingly, the bin k corresponding to a centre frequency f is given as

$$k(f) = \text{round}\left(\frac{f \cdot m \cdot N}{f_s}\right) \tag{1}$$

To characterise the presence of the voice, we compute the spectral band ratio $S[n]$ from the ratio of summed magnitudes in the upper band $f_{21} = 500\,\text{Hz} < f < f_{22} = 6000\,\text{Hz}$ to the lower band $f_{11} = 80\,\text{Hz} < f < f_{12} = 400\,\text{Hz}$:

$$S[n] = 20 \cdot \log 10 \left(\frac{\sum\limits_{k(f_{21})<k<k(f_{22})} |\dot{X}[k,n]|}{\sum\limits_{k(f_{11})<k<k(f_{12})} |\dot{X}[k,n]|}\right) \tag{2}$$

In order to eliminate the influence of the overall volume, we divide the magnitude spectrum $|X[k,n]|$ by its maximum value in the given frame $X_{\max}[n] = \max\limits_{k}(X[k,n])$, resulting in the normalised magnitude spectrum $|\dot{X}[k,n]|$ at time frame n. We compute the spectral band ratio $S[n]$ for each frame n and average for each channel separately over the entire length of the song. Based on the resulting two average spectral band ratios, $\overline{S_{\text{left}}}$ and $\overline{S_{\text{right}}}$, we select the channel with the higher average value. Preliminary experiments have shown that comparing the summed bin magnitude values $|X[k,n]|$ yields to a better discrimination than comparing the summed signal energy $|X[k,n]|^2$ among the two bands.

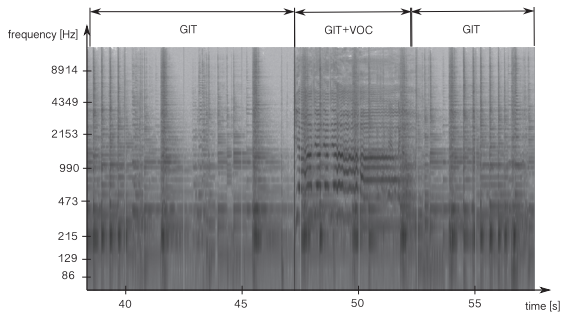

Fig. 2. Spectrogram for voiced and unvoiced sections: "GIT" denotes the presence of the guitar; "VOC" denotes the presence of the singing voice.

2) Predominant Melody Extraction: The predominant melody extraction algorithm [26] estimates pitch candidates on a frame level based on harmonic summation and groups them into pitch and time continuous contours using auditory streaming principles. It furthermore filters out contours based on their average pitch salience which do not form part of the dominant melodic line. As a result, it outputs a single pitch value $f_0[n]$ in Hz for all melody frames. As this algorithm has previously been used in the context of flamenco singing, we adopt the parameters suggested in [25] as follows: The analysis window corresponds to $N = 4096$ samples with a hop size of $h_s = 128$ samples at a sample rate of $f_s = 44.1$ kHz. The lower and upper limits for considered fundamental frequency candidates are set to 120 Hz and 720 Hz, respectively. These values correspond to the expected pitch range in flamenco singing. The voicing threshold $\tau_v[-2,3]$ which is related to the relative salience threshold for determining if a contour belongs to the main melody, is adjusted to $\tau_v = 0.2$. It has been shown [25] that with this value, around 90% of all vocal frames are retained during the elimination process.

3) Contour Filtering: First, we extract the energy in the lower twelve bark bands with frequency limits specified as 0 Hz, 50 Hz, 100 Hz, 150 Hz, 200 Hz, 300 Hz, 400 Hz, 510 Hz, 630 Hz, 770 Hz, 920 Hz, 1080 Hz and 1270 Hz in windows of length $N = 1024$ samples with a hop size of $h_s = 128$ samples at a sampling rate of $f_s = 44.1$ kHz. The energy in the m^{th} bark band with lower frequency limit $f_{1,m}$ and upper frequency limit $f_{2,m}$ at time frame n is given as

$$B[n,m] = \sum_{k(f_{1,m})<k<k(f_{2,m})} |X[k,n]|^2 \tag{3}$$

where $k(f)$ is the frequency bin corresponding to the frequency f (Eq. (1) and $|X[k,n]|$ is the magnitude spectrum at time frame n. As a result, we obtain a feature vector

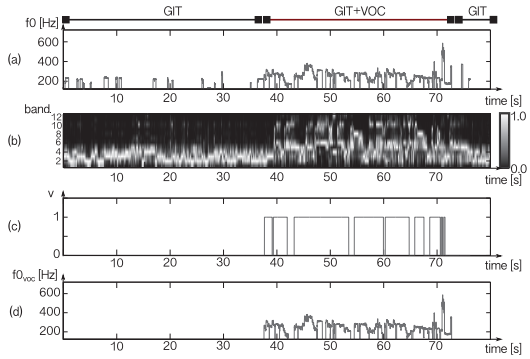

Fig. 3. Contour filtering: (a) predominant melody, (b) lower bark bands, (c) vocal / non-vocal classifier and (d) estimated vocal melody.

$\mathbf{x}[n] = \langle B[n,1], \ldots, B[n,12] \rangle$ holding the energies of the lower twelve bark bands for each analysed frame n.

We then assign an initial label to the feature vector in each frame based on the output of the predominant melody algorithm: Melody frames are labelled as *voiced* $\mathbf{x}_+$ and non-melody frames as *unvoiced* $\mathbf{x}_-$. This initial labelling corresponds to the case where the main melody coincides with the vocal melody. Subsequently, we fit a single multivariate Gaussian distribution to both feature sets separately. Applying maximum likelihood estimation, we obtain estimates for mean ($\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$) and covariance ($\boldsymbol{\Sigma}_+$ and $\boldsymbol{\Sigma}_-$) for both classes. The resulting likelihood p for an arbitrary feature vector $\mathbf{x}$ is given as:

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^{12}|\boldsymbol{\Sigma}|}} \cdot e^{(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))} \tag{4}$$

Evaluating this equation for a given feature vector $\mathbf{x}$ for both distributions, voiced $p_+(\mathbf{x}|\boldsymbol{\mu}_+, \boldsymbol{\Sigma}_+)$ and unvoiced $p_-(\mathbf{x}|\boldsymbol{\mu}_-, \boldsymbol{\Sigma}_-)$, we now expect a higher value for p_+ if the frame contains vocals. In this manner, we evaluate every frame and assign a binary vocal prediction $v[n]$:

$$\nu[n] = \begin{cases} 1 \text{ if } p_+(\mathbf{x}[n]|\boldsymbol{\mu}_+, \boldsymbol{\Sigma}_+) \geq p_-(\mathbf{x}[n]|\boldsymbol{\mu}_-, \boldsymbol{\Sigma}_-) \\ 0 \text{ else} \end{cases} \tag{5}$$

Since we assume voiced section to be continuous in time, we subsequently apply a binary moving average filter of 1 second length to the vocal prediction sequence $v[n]$ to eliminate fast fluctuations. The resulting sequence is shown in Figure 3 (c).

We now use this frame-wise classification to filter contour segments. A given contour ranging from frame d_1 to frame d_2 is eliminated, if it is entirely located outside the estimated vocal regions:

$$\sum_{d_1}^{d_2} v[n] \begin{cases} = 0 \text{ eliminate} \\ > 0 \text{ retain} \end{cases} \tag{6}$$

An example of the resulting vocal pitch is depicted in Figure 3 (d).

B. Note Transcription

After eliminating unwanted guitar contours, we are now left with a set of contours corresponding to the vocal melody. Instead of a continuous time-frequency representation, we now aim to segment the remaining contours into discrete note events, characterised by onset time, duration and a quantised semi-tone pitch value. In other words, the task is to split the contours at vocal onsets and assign a pitch label to each note event.

1) Segmentation: Conceptually, we can distinguish two types of notes onsets: Those which coincide with a change in pitch (*interval onsets*) and onsets where the pitch of the current note is the same as the previous note (*steady pitch onsets*). In what follows, we will define four detection schemes which in their combination capture the majority of both types of onsets at a low false positive rate.

For each voiced segment, we first map the frequency contour $f_0[n]$ to a cent scale relative to a reference frequency of $f_{\text{Ref}} = 440\,\text{Hz}$:

$$c_0[n] = 1200 * \log_2 \left(\frac{f_0[n]}{f_{\text{Ref}}} \right) \tag{7}$$

First, we aim to discover *interval onsets*: When dealing with strongly ornamented pitch contours, the detection of significant pitch changes which indicate a note onset turns into a complex task. For the particular case of flamenco singing, fast fluctuations of the pitch track caused by micro-tonal ornamentations can exceed a semitone without a note onset being present. Low-pass filtering of the pitch contour reduces these fluctuations but also affects the steep slopes which indicate a note change. A close inspection of such pitch contours has shown that the relative change of the upper envelope gives a good indication of the underlying slowly changing perceived pitch (Figure 4): Vocal vibrato causes a series of local maxima and in case of a steady note their pitch values remain in a strongly limited range in contrast to the actual pitch contour which might show fluctuations with an extend up to a semitone. In preliminary experiments the upper envelope has proven to be slightly more reliable for this task than the lower envelope. We furthermore observe that due to the periodicity of the vibrato fluctuations, the steep slope characterising a note onset is centred between two adjacent local maxima. We therefore extract the local maxima pm and assume a pitch change centred between two adjacent maxima if their relative distance $\Delta\text{pm}_{i,i+1}$ exceeds a threshold Δp_{min}. For a pitch change of an exact semitone, we expect $\Delta\text{pm}_{i,i+1} = 100$ cents. In order to leave room for intonation inaccuracies, we adjust the threshold to $\Delta p_{\text{min}} = 80\,\text{cents}$. Since not all local maxima are caused by vocal vibrato, we furthermore exclude cases where the time distance between two local maxima exceeds $T = 0.25\,\text{s}$. We chose this values in order to cover modulation rates as low as $4\,\text{Hz}$, which corresponds to the lower bound of expected vocal vibrato rates [35]. With this procedure we detect a large number of pitch interval onsets, in particular in sections where vocal vibrato is present. Nevertheless, a number of onsets are missed, i.e. when the contour does not show a regular vibrato and consequently no relevant local maxima in the area of a pitch change.

In order to detect further undiscovered onsets, we refine the segmentation by applying a first derivative Gaussian filter to the contour. Such filters have been used successfully in the related task of edge detection in image processing [33], [34]. The one-dimensional Gaussian density function $g[n, \sigma]$ with zero mean

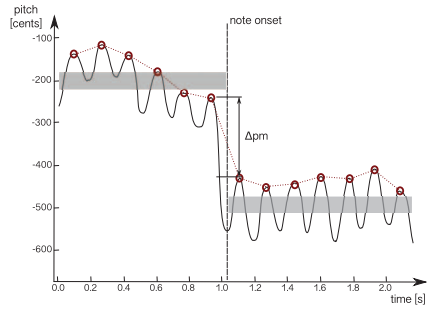

Fig. 4. Note segmentation based on local maxima: Ground truth transcription (grey rectangles), pitch contour, local maxima (red circles) and upper envelope (red dashed line).

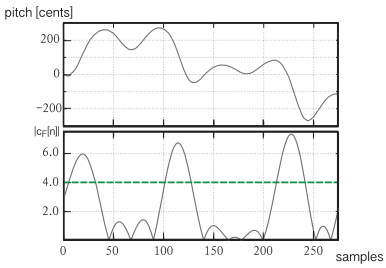

Fig. 5. Gaussian derivative filtering; top: cent-scaled pitch contour; bottom: output of the Gaussian derivative filter. The green dashed line marks the detection threshold for local maxima.

and standard deviation σ is given as:

$$g[n,\sigma] = e^{-\frac{n^2}{2\sigma^2}} \tag{8}$$

and its first derivative $h[n]$ results to:

$$h[n] = \frac{\delta g[n,\sigma]}{\delta n} = -\frac{n}{\sigma^2} \cdot e^{-\frac{n^2}{2\sigma^2}} \tag{9}$$

In the context of onset detection, the purpose of applying Gaussian derivative filtering is to detect long-term changes in the signal while omitting fast fluctuations. The parameter σ determines the effective length of the filter and consequently the period of averaging: If σ is too large, the filter might average entire short notes. Choosing a very small value for σ will lead to a large filter output at fast pitch fluctuations caused by vibrato. We choose $\sigma = 43.5$ ms, so that the effective filter length of 300 ms covers a full period of a slow vibrato with a rate of 4 Hz. Applying the filter $h[n]$ of length N_h to the cent-scaled contour $c_0[n]$ of length N_c leads to the filtered output $c_F[n]$. Analysing its absolute value (Figure 5) shows strong peaks in the area where a change of the slowly varying underlying pitch takes place. We consequently detect onsets at local maxima of $|c_F[n]|$. Since minor variations in the pitch contour may cause noise in the filter output, we discard local maxima below an empirically determined threshold of $|c_F[n]|_{\text{min}} = 4.0$.

We now proceed to the detection of *steady pitch onsets*: We define two characteristics which indicate this type of onset: A sudden local decay in volume (Figure 6) and sudden decrease in the pitch contour (Figure 7). It is important to state at this point, that at a given onset either one or both of these characteristics can be present. We therefore define two separate

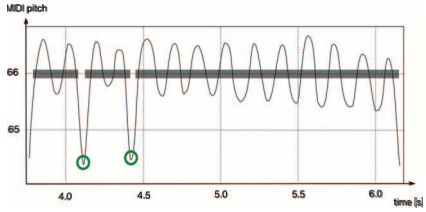

Fig. 6. Pitch contour dips at note onsets: pitch contour; note onsets characterised by a dip in the pitch contour. Green circle mark detected onsets, grey rectangles correspond to the ground truth transcription.

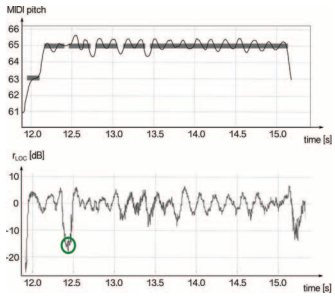

Fig. 7. Note onsets characterised by a local decrease in volume; top: pitch contour; bottom: local RMS; The green circle marks a detected onset, grey rectangles correspond to the ground truth transcription.

detection schemes instead of combining both features in a single detection function. Detecting volume decays when the overall dynamics of a track vary, i.e. certain sections are generally of lower volume than others, requires an analysis of the short-term dynamics. We first extract the root mean square (RMS) of the signal $\text{rms}[n]$ in windows of length $N = 4096$ samples with a hop size of $h_S = 128$ samples. In order to detect local decays, we define the local RMS fluctuation $r_{\text{LOC}}[n]$ by comparing each sample to the mean value of its surrounding 100 samples, corresponding to 150 ms and map to a decibel scale:

$$r_{\text{LOC}}[n] = 20 \cdot \log 10 \left(\frac{\text{rms}[n]}{\sum\limits_{-50 \le n \le 50 \wedge n \neq 0} \text{rms}[n] \cdot 0.01} \right) \tag{10}$$

We segment the contour at frames n where the local minima with $r_{\text{LOC}}[n] < -10$ dB are found. We set this threshold in order to ignore local volume variations which often accompany pitch vibrato or are caused by dynamic fluctuations of the accompaniment.

The difficulty in detecting the previously described decreases in the pitch envelope is to distinguish them from local minima occurring during vocal vibrato. We therefore model local minima which are related to note onsets as outliers in the local distribution of pitch values over all frames in the considered contour. A common method to detect outlier values of a distribution is to analyse its z-score $z[n]$, which describes the relation between a given datapoint and the standard deviation of all considered data points,

$$z[n] = \frac{c[n] - \mu}{\sigma} \tag{11}$$

where μ denotes the mean and σ the standard deviation of the distribution of pitch values in the considered segment. Consequently, the described decreases in a given contour will produce large negative values for $z[n]$. In order to avoid detecting local minima caused by vocal vibrato, we only consider negative peaks of $z[n]$ below a threshold $z_{\max}$. The standard deviation of a zero-centred sinusoid $\sigma_{\sin}$ with a magnitude A corresponds to its RMS value and results to:

$$\sigma_{sin} = \frac{A}{\sqrt{2}} \tag{12}$$

Local minima of the periodic oscillation will consequently cause a z-score $z_{\min,\sin}$ of

$$z_{\min,\sin} = \frac{-A}{\sigma_{\sin}} = -\sqrt{2} \tag{13}$$

Since we are looking for local pitch decreases which ares significantly larger than the deviations caused by vocal vibrato and furthermore the vibrato extend might vary during long notes, we adjust the threshold for detecting local minima below the theoretical boundary of $-\sqrt{2}$ to $z_{\max} = -2$.

All previously described threshold values involved in the onset detection process ($|c_F[n]|_{\min} = 4.0$, $r_{\mathrm{LOC}}[n] < -10\,\mathrm{dB}$ and $z_{\max} = -2$) were determined empirically based on observations of several onsets showing the respective characteristics which were singled out from the *cante2midi* dataset (Section III-A).

2) Pitch Labelling: After the note segmentation stage, the remaining task is to assign a pitch label to each resulting note event. There are various challenges in this stage: First, the tuning of the track might deviate from the reference, causing even constant segments to be located between two semitone bins. While the mean pitch value across the contour is often a sufficient approximation for stable contour segments or even symmetric ornamentations (i.e. vibrato), singing voice contours may contain portamentos at note beginnings and endings or over-swing before the target pitch is reached. Such non-symmetric ornamentations can cause an offset in the mean pitch value. Finally, local intonation might not always be accurate and the estimated pitch can be located slightly below or above the target value. McNab [16] estimates the resulting pitch as the peak bin of the local pitch histogram. Molina *et al.* determine the pitch label of a given segment as the energy-weighted alpha-trimmed mean in order to exclude local pitch outliers and give assign higher weights to high-energy frames. In the model-based transcription system [18], the pitch label of each found note results directly from the output of the HMM. In a similar way, Gómez & Bonada [23] deduce quantised pitch labels from the maximum likelihood estimation.

Here, we first estimate a global tuning reference for the entire song and then combine pitch statistics from the note event with global pitch class probabilities to finally determine the pitch value. The entire process is summarised in Figure 8.

We estimate a global tuning deviation Δt from the reference value of $A_{4,\mathrm{St}} = 440\,\mathrm{Hz}$ using an established method based on circular statistics [36]. The new tuning reference $A_{4,T}$ results to

$$A_{4,T} = 2^{\frac{\Delta t}{1200}} \cdot A_{4,\mathrm{St}} \tag{14}$$

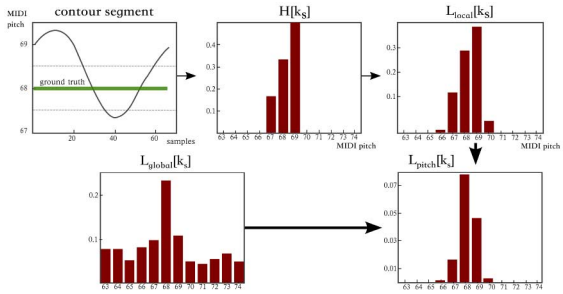

Fig. 8. Pitch labelling: (a) contour segment; (b) local pitch histogram; (c) local pitch probability; (d) global pitch probability; (e) combined pitch probability.

and we can re-map each contour $f_0[n]$ to a cent scale $c_{0,T}[n]$ relative to the estimated reference frequency $f_{\mathrm{Ref}} = A_{4,T}$ applying Eq. 7.

Even though in several music traditions and genres, the main melody does not strictly follow a defined scale, certain pitches tend to occur more frequent than others according to the underlying harmonic context. For note events with unstable pitch or inaccurate intonation, such probabilities can assist in deciding on the pitch label. We estimate a so called pitch class profile providing probability estimates for all twelve semitones from the averaged chroma vector over all frames. Chroma features were first introduced by Wakefield [37] and are frequently used in the context of a variety of MIR tasks, such chord and tonality estimation or cover song identification [38]. The chroma vector for a given time frame is obtained by quantising the spectrum $|X[n,k]|$ to semi-tone bins and then mapping the entire analysed range into a single octave. In this manner, we obtain an instantaneous chroma vector $\mathrm{chr}[k_{\mathrm{chr}}, n]$ for each frame n consisting of $K = 12$ semitone bins k_{chr}. Subsequently, we can estimate a global pitch class probability L_{global} for each semitone from the average chroma vector $\overline{\mathrm{chr}[k_{\mathrm{chr}}]}$ over all frames in a given track:

$$L_{\mathrm{global}}[k_{\mathrm{chr}}] = \frac{\overline{\mathrm{chr}[k_{\mathrm{chr}}]}}{\sum_{k_{\mathrm{chr}}=1}^{12} \overline{\mathrm{chr}[k_{\mathrm{chr}}]}} \tag{15}$$

The use of chroma vectors give the advantage that it includes pitch information of both, the singing voice and the guitar accompaniment, and should therefore give a better representation of the underlying tonality.

We now proceed to the statistical analysis of the pitch content within a single note event. The centre values $C_c[k_s]$ in cents of quantised cent bins corresponding to the k_s^{th} semitone above $A_{4,T}$ are given as:

$$C_c[k_s] = k_s \cdot 100 \tag{16}$$

Now, for each frame within the note event, the pitch contour $c_{0,T}[n]$ is quantised to the closest semitone bin k_s and accordingly a pitch histogram $H[k_s]$ is computed by accumulating the occurrences of each bin over all frames of the analysed track and dividing by the total number of frames. As mentioned earlier, due to local intonation inaccuracies and non-symmetric ornamentations, the ground truth pitch might not correspond

to the peak bin but to one of the adjacent bins. We therefore replace each bin value $H[k_s]$ by a Gaussian distribution $G_{k'_s}[k_s]$ originating from the bin k'_s and spanning over various semitone bins k_s:

$$G_{k'_s}[k_s] = H[k'_s] \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(k_s - k'_s)^2}{2\sigma^2}} \tag{17}$$

The mean of the distribution corresponds to the semitone bin k'_s from which the distribution originates. We furthermore assume a standard deviation of $\sigma = 0.5$ corresponding to a quarter tone and weight with the occurrence of the respective bin $H[k'_s]$. By accumulating the contributions of all distributions for each bin, we obtain the local pitch probability function $L_{local}[k_s]$ as a Gaussian mixture distribution:

$$L_{\text{local}}[k_s] = \frac{\sum_{k'_s} G_{k'_s}[k_s]}{K} \tag{18}$$

The pitch label can now be estimated by combining the global pitch class probability and the local pitch probability function. For a given semitone bin k_s the corresponding chroma bin k_{chr} is defined as:

$$k_{\text{chr}}[k_s] = k_s \bmod 12 \tag{19}$$

We calculate the product of global and local pitch class probabilities $L_{\text{pitch}}[k_s]$ as:

$$L_{\text{pitch}}[k_s] = L_{\text{global}}[k_{\text{chr}}[k_s]] \cdot L_{\text{local}}[k_s] \tag{20}$$

The MIDI pitch label $P_{\text{midi},i}$ associated with the i^{th} contour segment is finally computed from the semitone bin with the highest combined pitch probability:

$$P_{\text{midi},i} = 69 + \underset{k_s}{\text{argmax}}(L_{\text{pitch},i}[k_s]) \tag{21}$$

3) Note Post-Processing: By applying basic musicologically motivated restrictions regarding pitch and duration, we can further refine the obtained note transcription. First, we can exploit the idea that flamenco singing is usually limited to a pitch range of less than an octave. Figure 9 shows an analysis of the *cante2midi* dataset described in Section III, where the pitch of a ground truth note is displayed in relation to the median pitch of the corresponding track. Based on this observations, we subtract the median pitch of the entire track from each transcribed note and limit the range to ± 8 semitones. Transcribed notes below this range are likely to belong to the accompaniment and are consequently eliminated. Notes above this range may be caused by octave errors in the vocal melody extraction and are transposed down by one octave. We furthermore assume a minimum note duration of 0.05 seconds and eliminate shorter transcribed notes. In the analysed datasets (III-A), only 0.3% of all ground truth notes have a shorter duration. In order to reduce computational complexity and to avoid labelling short notes which are afterwards discarded, segments shorter than 0.05 seconds are eliminated before the note labelling stage.

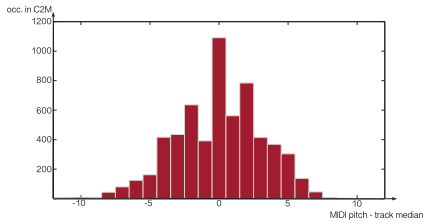

Fig. 9. Distribution of ground truth pitch values with respect to the track median for the *cante2midi* dataset.

TABLE I
DATABASE INFORMATION

database	cante2midi	fandangos	tonas
no. tracks	20	39	72
clip type	full track	excerpt	excerpt
no. singers	15	21	44
total duration	1h 6m	34m	20m
no. ground truth notes	6025	3070	2983
% voiced frames	42	50	82

III. EVALUATION METHODOLOGY

Below we provide a detailed description of the evaluation strategies applied in the scope of this study. We give an overview of the employed data collections and the corresponding ground truth annotation process and describe the evaluation metrics used. We furthermore provide a short description of the reference algorithms used during the comparative evaluation.

A. Test Collections

We use three data collections to evaluate the performance of the proposed approach and compare to reference algorithms: The *cante2midi* (C2M) dataset was gathered in the scope of this study and contains a variety of flamenco styles and voice timbres with varying degree and complexity of ornamentation. The instrumentation of all 20 tracks is limited to vocals and guitar. The *fandango* (FAN) dataset comprises 39 excerpts from recordings of the fandango style, containing vocals and guitar accompaniment. This dataset was used previously in the context of automatic transcription of flamenco singing [25]. In order to compare to monophonic transcription systems, we furthermore evaluated the proposed system on the *tonas* (TON) dataset. This collection contains 72 clips of a cappella flamenco recordings and is publicly available[2]. It should be mentioned that a cappella singing represents only a small fraction of the flamenco genre and such performances are usually characterised by a large amount of melismatic ornamentation. Furthermore, the absence of the guitar accompaniment often causes strong tuning fluctuations throughout the performance. Further information on all three data collections is provided in Table I.

B. Ground Truth Annotations

A general guideline for the ground truth annotation process of all three collections was to obtain a detailed transcription

[2]mtg.upf.edu/download/datasets/tonas/

including all audible notes. Since flamenco music does not always follow a strict rhythm and the proposed transcription system does not apply rhythmic quantisation, note onsets were transcribed as absolute time instants. The ground truth annotations for the C2M collection were conducted by a person with formal music education and training in melody transcription by ear, but only basic knowledge of flamenco. All transcriptions were verified and corrected by a flamenco expert. The output of the system described by Gómez *et al.* [25] was taken as a starting point. After converting to MIDI, transcriptions were edited in the digital audio workstation *LogicPro*. The annotator listened to the original track and the transcription synthesised by a piano simultaneously with the possibility of muting one of the tracks when required. The tuning of the MIDI synth was manually adjusted to match the tuning of the corresponding audio track. A visual representation of the pitch contour and the baseline transcription was provided as additional aid. Ground truth annotations for both, the FAN and the TON collection, were conducted by a musician with limited knowledge of flamenco in order to avoid implicit knowledge of the style. Annotations were then verified and corrected by a flamenco expert and occasionally discussed with another flamenco expert. For more details on the annotation process of these two collections and general guidelines of manual flamenco singing transcription, we refer to [23] and [25]. In addition to the note transcriptions, we furthermore provide manually corrected pitch contours for both polyphonic databases, where guitar contours were eliminated.

C. Reference Algorithms

In the course of this study we compared the proposed system to a number of reference algorithms. Below we briefly describe each of the methods we used in the comparative evaluation in Section IV. For a detailed description we refer the reader to the references provided for each method.

- *Curve fitting (FIT)*: A contour simplification algorithm [2] was applied to a given pitch contour. As suggested by the authors, the pitch deviation tolerance was set to α_e=1.
- *Recursive least squares filtering (RLS)*: This approach performs a segmentation of the pitch contour based on the error function of an adaptive filter which tracks the semitone pitch contour as described by Adams [17]. We adopt all system parameters suggested by the authors and segment at values of the squared error function $d[n] > 0.25$.
- *Dynamic programming (DP)*: We used an existing implementation of the flamenco singing transcription system described by Gómez & Bonada [23] (DP-Mono) to transcribe a cappella recordings and applied its extension [25] (DP-Poly) for polyphonic recordings. The implementation furthermore allows to segment any given pitch sequence.
- *Segmentation based on hysteresis (SiPTH)*: The monophonic singing transcription described by Molina *et al.* [21] was used to transcribe a cappella singing recordings.
- *Segmentation based on probabilistic Hidden Markov Modelling (HMM)*: We compare to the implementation of the segmentation algorithm proposed in [18] as implemented in the *tony* [20] transcription framework for monophonic singing recordings. The system uses the probabilistic YIN [39] algorithm as a front end to estimate the vocal pitch. We used the publicly available vamp plugin implementation[3].

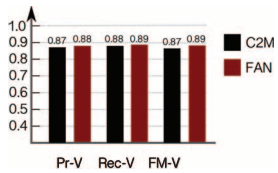

Fig. 10. Frame-wise accuracy of the GMM-based vocal detector [28]: *cante2midi* and *fandango* dataset.

The systems SiPTH and HMM represent monophonic singing transcription frameworks and were consequently evaluated in the scope of a cappella singing. FIT and RLS are contour segmentation algorithms without a vocal pitch extraction front end and were used in the scope of onset detection evaluation. The DP algorithm is the only complete reference system for singing transcription from polyphonic recordings and was therefore part of a comparative study evaluating the overall system performance. Since it can furthermore operate on any given pitch contour input, it was also used in the onset detection evaluation.

- *Vocal detection using Gaussion Mixture Models (GMM)*: In order to evaluate the proposed vocal detection stage to existing methods, we furthermore implemented a the algorithm proposed by Song *et al.* [28] based on Gaussian Mixture Models. We adopted all parameters as suggested by the authors and processed both polyphonic datasets in a 10-fold cross-validation. The frame-wise accuracy by means of voicing precision, recall and f-measure is shown in 10 and is slightly above the values reported by the authors

D. Evaluation Metrics

We aim to evaluate three aspects of the proposed algorithm: Its capability of detecting the vocal sections, the performance of the segmentation of vocal contours into discrete note events and its overall performance.

The vocal section retrieval is evaluated by means of voicing precision (Pr-V) , voicing recall (Rec-V) and voicing f-measure (FM-V) . Pr-V is defined as the fraction of all frames estimated as voiced, which are labelled as voiced in the ground truth. Rec-V corresponds to the fraction of all voiced ground truth frames, which are estimated as voiced. The resulting f-measure is calculated as the harmonic mean of Rec-V and Pr-V.

In a similar manner, we evaluate the onset detection stage by means of onset precision (Pr-On), onset recall (Rec-On) and onset f-measure (FM-On). In this case, Pr-On refers to the proportion of all detected onsets, which correspond to ground truth onsets. Rec-On is defined as the proportion of all ground truth onsets, which are correctly detected. FM-On again corresponds to the harmonic mean of Rec-On and Pr-On. We adapt

[3]code.soundsoftware.ac.uk/projects/pyin

a previously suggested threshold ([23]) and consider an onset as correctly detected if it is located within 0.15 seconds of a ground truth onset. Furthermore, each ground truth onset can only be associated to a single detected onset and vice versa.

In order to evaluate note transcriptions, we first define the frame-wise raw pitch accuracy (RPA) as the percentage of correctly transcribed frames. In order to incorporate the voicing detection into this measure, we define an unvoiced frame as correctly transcribed if it was estimated as unvoiced. Voiced frames are correctly transcribed if the estimated MIDI pitch corresponds to the ground truth in this frame. We furthermore evaluate transcriptions based on note precision (Pr-N) , note recall (Rec-N) and note f-measure (FM-N) . Pr-N refers to the proportion of all detected notes, which are correctly transcribed ground truth notes. Rec-N is defined as the proportion of all ground truth notes, which are correctly transcribed and FM-N is calculated as the harmonic mean of Pr-N and Rec-N.

We adopt the evaluation thresholds suggested in [23]: A ground truth onset is correctly detected if an estimated onset is within a range of 0.15 seconds. Furthermore, a note is correctly transcribed if the pitch label is correctly assigned and the estimated duration is within a tolerance of 30%. Analysing the test collection, we observe significant deviations from the standard tuning of $A_4 =$ Hz ranging even above 40 cents. For such large deviations, it is even in the manual process difficult to decide, if the track is tuned below or above the reference. Consequently, small errors in the tuning estimation of only a few cents can cause the entire melody to be transcribed a semitone above or below the ground truth. We assume that in cases of large tuning deviations such a transcription should still be valid. Therefore, we perform a preliminary evaluation of the entire transcription and its transpositions one semitone above and below and correct towards the best match.

IV. Experimental Results

In this section, we present the results of a number of experiments carried out in order to evaluate the performance of the proposed system (P) and compare to other methods. We first evaluate the overall performance for flamenco transcription from polyphonic and monophonic systems and compare to a several existing systems. In order to deal with monophonic recordings, we omit the channel selection and contour filtering stage and increase the voicing tolerance of the predominant melody extraction algorithm (Section II-A2) to $\tau_v = 3.0$. This setup is denoted as P-Mono. We then analyse the accuracy of the proposed contour filtering stage, compare to alternative system setups and investigate the influence on the note transcription. Subsequently, we first isolate the contour segmentation stage and then the entire note transcription block and study the obtained accuracies with respect to alternative approaches. Finally, we conduct a component analysis of the proposed system and investigate the influence of certain processing blocks on the overall performance.

A. Overall System

Figure 11 shows the evaluation of the proposed system (P) and the polyphonic implementation of the dynamic programming approach (DP-Poly) [25] for the two polyphonic datasets described in Section III-C. Figure 12 shows the comparative evaluation on the monophonic dataset. For both, monophonic and polyphonic recordings, the proposed system outperforms all reference systems. The note f-measure as well as the frame-wise raw pitch accuracy is significantly lower for the monophonic dataset, indicating the difficulty of transcribing this particular sub-genre (Section III-A).

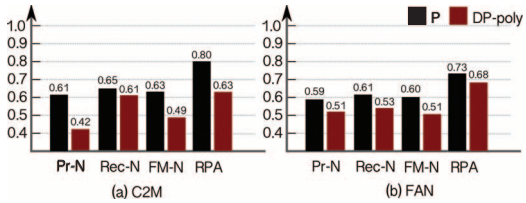

Fig. 11. Overall system evaluation for polyphonic datasets: Proposed system (P) and [25] (DP-Poly).

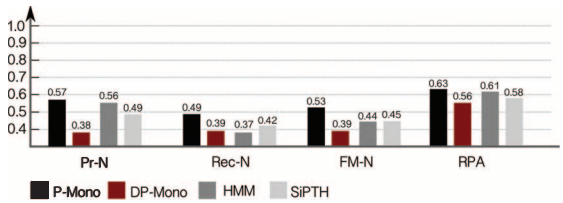

Fig. 12. Overall system evaluation for the monophonic dataset (TON): Proposed system (P), [23] (DP-Mono), [18], [20] (HMM) and [21] (SiPTH).

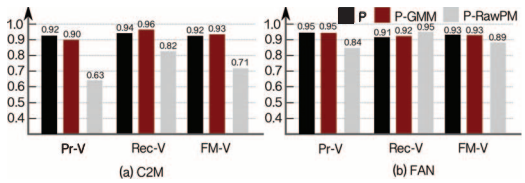

Fig. 13. Frame-wise voicing accuracy for the proposed system (P), the proposed system with [28] replacing the vocal detection function $v[n]$ (P-GMM) and the raw predominant melody (P-RawPM).

B. Voicing

We now investigate the effectiveness of the proposed vocal pitch extraction stage (P) and compare to two alternative setups: P-RawPM refers to replacing the entire stage with the raw predominant melody without any further processing. This setup corresponds to the front-end of the algorithm proposed by Gómez & Bonada [25] and represents the baseline. We furthermore replace the proposed frame-wise voicing prediction $v[n]$ (Section II-A3) with the output of the GMM-based approach described in [28] and eliminate contours accordingly. This setup is referred to as P-GMM. The experiments were conducted for both polyphonic datasets, evaluating the frame-wise voicing accuracy (Figure 13) as well as the influence on the resulting note transcription (Figure 14).

The results show that the raw predominant melody gives significantly better results for the *fandango* (FAN) than for the *cante2midi* (CSM) dataset. An explanation for this behaviour can be found when analysing the content of the excerpts comprising FAN: Despite containing only 50% of voiced

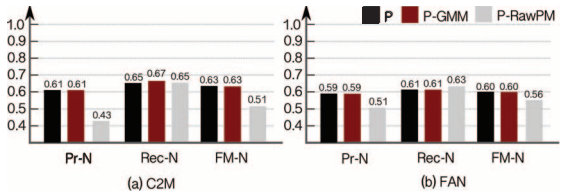

Fig. 14. Note transcription evaluation for the proposed system (P), the proposed system with [28] replacing the vocal detection function $v[n]$ (P-GMM) and the raw predominant melody (P-RawPM).

frames (Table I), the excerpts do not include the guitar introduction or melodic interludes. In contrast, C2M contains full tracks where these sections are included. These parts of the song represent the sections, where the guitar is most dominant and consequently tends to produce melody contours. Nevertheless, applying channel selection and contour filtering leads to a significant increase in the obtained frame-wise voicing accuracy which propagates to the resulting transcription quality: The proposed approach yields a note f-measure of 0.63 on the C2M dataset and 0.60 on the FAN dataset. The note f-measure obtained with the raw predominant decreases 0.51 for C2M and 0.56 for FAN. We can furthermore observe, that replacing $v[n]$ with the output of the GMM-based vocal detection yields similar results when compared to the proposed system. The difference in voicing accuracy between the two setups is marginal and does not show in the resulting not transcription performance (both obtain an f-measure of 0.63 on C2M and 0.60 on FAN). Nevertheless, an advantage of the proposed system is that it works on a track-level and does not require any training phase involving manual ground truth annotations.

C. Segmentation

We now isolate the note segmentation stage described in Section II-B1 and compare to a number of alternative approaches. The evaluation is carried out by means of onset precision, recall and f-measure. For all three databases, the corrected vocal pitch contour is provided as input to all methods. The results shown in Figure 15 show that the proposed approach (P-SEG) yields the best performance among all considered methods on all datasets, followed by the dynamic programming approach (DP-SEG) [23]. The curve fitting algorithm (FIT-SEG) [2] obtains a low precision but a high recall rate, indicating an over-segmentation of the contour. The reversed behaviour is observed for the RLS segmented (RLS-SEG) [17]. We furthermore observe, that the performance of the proposed approach is consistent for the three datasets. The onset f-measure of DP-SEG decreases for the monophonic dataset (0.78 for C2M, 0.76 for FAN and 0.68 for TON). Given the higher complexity of the segmentation task for the *tonas* dataset described in Section III-A, these results indicate that the proposed approach is robust towards tuning inaccuracies and is capable of dealing with a large amount of melismatic ornamentations of the melody.

D. Note Transcription

After isolating the note segmentation stage, we now proceed to the evaluation of the entire note transcription

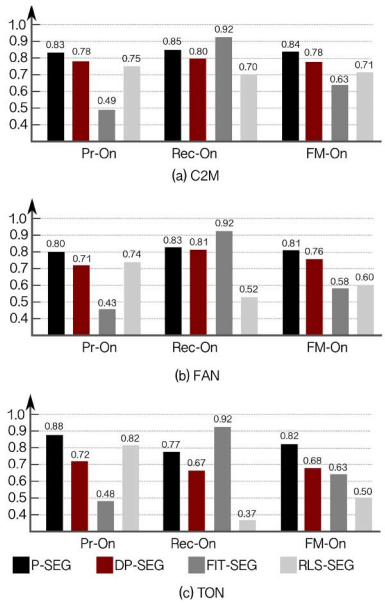

Fig. 15. Note segmentation evaluation for the proposed system (P-SEG), the dynamic programming approach (DP-SEG) [23], the fitting algorithm (FIT-SEG) [2] and the RLS segmenter (RLS) [17].

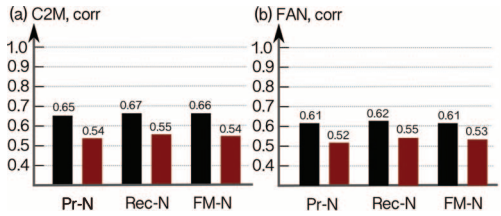

Fig. 16. Note transcription evaluation for the proposed system (P), the dynamic programming approach (DP) [23] when applied to the manually corrected pitch contour (*corr*).

stage (Section II-B), comprising contour segmentation, pitch labelling and note post-processing. We compare to the DP algorithm [23] and provide both systems with the manually corrected vocal pitch contour as input. This experimental setup represents a glass ceiling evaluation in a sense that it corresponds to a transcription with a perfect vocal pitch extraction stage in terms of voicing. Figure 16 provides the note-related evaluation for both polyphonic datasets. It can be observed that the decrease in performance from using the corrected pitch contours in this experiment to the real-world scenario (Figure 11) is significantly lower for the proposed system: For the C2M dataset, the note f-measure drops from 0.66 to 0.63 for the proposed system and from 0.54 to 0.39 for the DP algorithm. This indicates that the proposed system provides a better approximation of the vocal pitch contour than the raw predominant melody [23]. These results confirm the findings in Section IV-B.

E. Component Analysis

In a last experiment, we conduct a component analysis of the proposed system in order to verify that each of the core algorithm components contribute to the overall system performance. For this analysis, we chose the C2M dataset since it contains

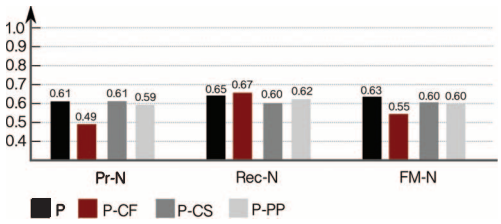

Fig. 17. Note transcription evaluation for the proposed system (P) when several components are removed: Contour filtering (P-CF), channel selection (P-CS) and global pitch probability estimation (P-PP).

recordings which are both polyphonic and stereo. We evaluate the proposed system (P) and compare to scenarios where single components are removed: The contour filtering stage (P-CF, Section II-A3), the channel selection (P-CS, Section II-A1) and the global pitch probability (P-PP, Section II-B2).

The results displayed in Figure 17 confirm that each of the investigated algorithm components contributes to the overall system performance. Among the considered setups, the removal of the contour filtering stage caused the largest decrease in performance, reducing the note f-measure from 0.63 to 0.55. Nevertheless, this result is still superior to the performance observed for the system presented in [25] (f-measure 0.49).

F. Summary

The results of the previously described experiments show that the proposed system gives superior results when compared to a number of reference algorithms. We observe a higher overall system performance when comparing to a state of the art flamenco singing transcription system (Figure 11). The note transcription f-measure for the proposed system ranges between 0.60 and 0.63 depending on the evaluation dataset. The results in Figure 13 show that the contour filtering process yields an improved voicing detection and gives similar results to a GMM-based vocal detection scheme which requires a pre-trained model (Figure 14). We isolated the note segmentation and pitch labelling stages and observed a superior performance compared to a number of state of the art systems (Figure 15 and 16). Finally, we conducted a component analysis and demonstrated that each processing block contributes to the overall system performance (Figure 17).

V. Conclusions

In this paper we present a novel approach for automatic transcription of flamenco singing directly from polyphonic audio recordings. All involved signal processing blocks have been described in detail: For stereo recordings we first select the channel with stronger dominance of the vocals for further processing. We then extract the predominant melody and apply a novel contour filtering process to discard guitar contours. It has been shown that this process significantly improves the voicing accuracy when compared to the raw predominant pitch contour. The resulting estimated vocal pitch sequence is further processed in the note transcription stage, where contours are converted to discrete note events. We propose a novel contour segmentation procedure based on pitch and volume characteristics, which has proven to yield better results by means of onset detection accuracy when compared to a number of alternative approaches. We assign a pitch label to each resulting segment by combining local and global pitch information. A number of experiments have shown that the overall system as well as isolated stages outperform various reference algorithms. Our approach has proven to give convincing results given with the particular characteristics of flamenco singing, in particular strong ornamentations, extensive use of vocal vibrato and local intonation inaccuracies. The resulting automatic transcriptions therefor provide a suitable basis for a number of related MIR tasks, such as melodic similarity characterisation or automatic style identification. The system can furthermore aid in large-scale musicological studies by providing first estimates for computer-assisted transcription.

There are several aspects in the context of flamenco singing transcriptions which we aim to investigate in order to further improve the quality of automatic transcription. A main topic of interest is to include perceptual aspects in the pitch labelling stage, by exploring the influence of fast pitch fluctuations and guitar accompaniment on the perceived pitch. Furthermore, it would be interesting to investigate how the quantitative measures presented in this study relate to perceptual transcription quality and how the transcription accuracy influences the performance of MIR systems which rely on automatic transcriptions. We are furthermore interested in how far more detailed note representations, i.e. including micro-tonal pitch labels compare the proposed standard MIDI representation for different MIR tasks. Finally, we aim to annotate ground truth for music traditions with similar characteristics to flamenco in order to evaluate the suitability of our approach for a larger variety of genres.

References

[1] F. Gómez-Martín, J. M. Díaz-Báñez, E. Gómez, and J. Mora, "Flamenco and its computational study," in *Proc. BRIDGES: Math. Connections Art Music Sci.*, Seoul, Korea, 2014, pp. 119–126.

[2] J. M. Díaz-Báñez and J. C. Rizo, "An efficient DTW-based approach for melodic similarity in flamenco singing," *Similarity Search and Applications, Lecture Notes in Computer Science*. New York, NY, USA: Springer, 2014, vol. 8821, pp. 289–300 [Online]. Available: http://link.springer.com/chapter/10.1007%2F978-3-319-11988-5_27

[3] N. Kroher and E. Gómez, "Computational models for perceived melodic similarity in a cappella flamenco cantes," in *Proc. 15th Int. Soc. Music Inf. Retrieval Conf.*, Taipei, Taiwan, 2014, pp. 65–70.

[4] N. Kroher and E. Gómez, "Automatic singer identification for improvisational styles based on vibrato, timbre and statistical performance descriptors" in *Proc. 11th Sound Music Comput. Conf.*, Athens, Greece, 2014, pp. 1160–1166.

[5] N. Kroher, A. Chaachoo, J. M. Díaz-Báñez, E. Gómez, F. Gómez-Martín, J. Mora, and M. Sordo, "Computational ethnomusicology: A study of flamenco and Arab-Andalusian vocal music," in *Handbook for Systematic Musicology*. New York, NY, USA: Springer, 2015.

[6] A. P. Klapuri, "Automatic music transcription as we know it today," *J. New Music Res.*, vol. 33, no. 3, 2004, pp. 269–282.

[7] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: Challenges and future directions," *J. Intell. Inf. Syst.*, vol. 41, no. 3, 2013, pp. 407–434.

[8] G. E. Poliner, D. P. W. Ellis, A. F. Ehrmann, E. Gómez, S. Streich, and B. Ong, "Melody transcription from music audio: Melody transcription from music audio: Approaches and evaluation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 4, pp. 1247–1256, May 2007.

[9] J. Salamon, E. Gómez, D. P. W. Ellis, and G. Richard, "Melody extraction from polyphonic music Signals: Approaches, applications and challenges," *IEEE Signal Process. Mag.*, vol. 31, no. 2, pp. 118–134, Mar. 2014.
[10] J. Mora, F. Gómez-Martín, E. Gómez, F. Escobar Borrego, and J. M. Díaz-Báñez, "Characterization and melodic similarity of a cappella flamenco cantes," in *Proc. Int. Symp. Music Inf. Retrieval*, Utrecht, Netherlands, 2010, pp. 351–356.
[11] J. M. Gamboa, *Una Historia Del Flamenco*. Madrid, Spain: Espasa-Calpe, 2005.
[12] J. A. Moorer, "On the transcription of musical sound by computer," *Comput. Music J.*, vol. 1, no. 4, pp. 32–38, 1977.
[13] M. Antonelli and A. Rizzo, "A correntropy-based voice to MIDI transcription algorithm," in *Proc. 10th IEEE Workshop Multimedia Signal Process.*, 2008, pp. 978–983.
[14] E. Pollastri,"A pitch tracking system dedicated to process singing voice for music retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, 2002 pp. 341-344.
[15] C.-K. Wang and R.-Y. Lyu, "A robust singing melody tracker using adaptive round semitones (ARS)," in *Proc. 3rd Int. Symp. Image Signal Process. Anal.*, 2003, pp. 549–554.
[16] R. J. McNab, L. A. Smith, and I. H. Witten, "Signal processing for melody transcription," in *Proc. 19th Aust. Comput. Sci. Conf.*, 1997, pp. 301–307.
[17] R. H. Adams, M. A. Bartsch, and G. H. Wakefield, "Note segmentation and quantisation for music information retrieval," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 1, pp. 131–141, Jan. 2006.
[18] M. Ryynänen and A. P. Klapuri, "Transcription of the singing melody in polyphonic music," in *Proc. 7th Int. Conf. Music Inf. Retrieval*, 2006, pp. 222–227.
[19] W. Krige, T. Herbst, and T. Niesler, "Explicit transition modelling for automatic singing transcription," *J. N. Music Res.*, vol. 34, no. 4, pp. 311–324, 2008.
[20] M. Mauch *et al.*, "Computer-aided melody note transcription using the tony software: Accuracy and efficiency," in *Proc. 1st Int. Conf. Technol. Music Notation Represent.*, 2015, pp. 25–30.
[21] E. Molina, L. J. Tardón, A. M. Barbancho, and I. Barbancho, "SiPTH: Singing transcription based on hysteresis defined on the pitch-time curve," *IEEE Trans. Audio Speech Lang. Process.*, vol. 23, no. 2, pp. 252–263, Feb. 2015.
[22] X. Serra, "A multicultural approach in music information research," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2011, pp. 151–156.
[23] E. Gómez and J. Bonada, "Towards computer-assisted flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to a cappella singing," *Comput. Music J.*, vol. 37, no. 2, pp. 73–90, 2013.
[24] J. J. Mestres, J. B. Sanjaume, M. De Boer, and A. L. Mira, "Audio recording analysis and rating," U.S. Patent 8,158,871, Apr. 17, 2012.
[25] E. Gómez, F. Cañadas, 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. 13th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2012, pp. 601–606.
[26] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 6, pp. 1759–1770, Aug. 2011.
[27] J. M. Díaz-Báñez and J. A. Mesa, "Fitting rectilinear polygonal curves to a set of points in the plane," *Eur. J. Oper. Res.*, vol. 130, no. 1, pp. 214–222, 2001.
[28] L. Song, M. Li, and Y. Yan, "Automatic vocal segments detection in popular music," in *Proc. 9th Int. Conf. Comput. Intell. Secur.*, 2013, pp. 349–352.
[29] M. Rocamora and P. Herrera, "Comparing audio descriptors for singing voice detection in music audio files," in *Proc. 11th Braz. Symp. Comput. Music*, 2007, pp. 187–196.
[30] S. D. You and Y.-C. Wu, "Comparative study of singing voice detection methods," in *Computer Science and its Applications, Lecture Notes in Electrical Engineering*. Heidelberg, Germany: Springer, 2015, vol. 330, pp. 1291–1298.
[31] D. Bogdanov *et al.*, "ESSENTIA: An open source library for audio analysis," *ACM SIGMM Rec.*, vol. 6, no. 1, pp. 493–498, 2014.
[32] E. Zwicker and E. Terhardt, "Analytical expressions for critical band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Amer.*, vol. 68, pp 1523–1525, 1980.
[33] M. Basu, "Gaussian-based edge-detection methods–A survey," *IEEE Trans. Syst. Man Cybern.*, vol. 32, no. 3, pp. 252–260, Aug. 2002.
[34] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.
[35] J. Sundberg, "Acoustic and psychoacoustic aspects of vocal vibrato," in *Vibrato*, P. Dejonckere, M. Hirano, and J. Sundberg, Eds. San Diego, CA, USA: Singular Publishing Company, 1995.
[36] K. Dressler and S. Strech, "Tuning frequency estimation using circular statistics," in *Proc. 8th Int. Conf. Music Inf. Retrieval*, 2007, pp. 357–360.
[37] G. H. Wakefield, "Mathematical representation of joint time-chroma distributions," in *Proc. Adv. Signal Process. Algorithms Architect. Implement. (SPIE)*, 1999, pp. 637–645.
[38] D. P. W. Ellis, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2007, pp. 1429–1432.
[39] M. Mauch and S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2014, pp. 659–663.
[40] N. Kroher, A. Pikrakis, J. Moreno, and J. M. Díaz-Báñez, "Discovery of repeated vocal patterns in polyphonic audio: A case study on flamenco music," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2015, pp. 41–45.
[41] E. Gómez, J. Bonada, and J. Salamon, "Automatic transcription of flamenco singing from monophonic and polyphonic music recordings," in *Proc. 3rd Interdiscip. Conf. flamenco Res. (INFLA)*, 2012, pp. 191–198.

Nadine Kroher received the M.Sc. degree in audio and electrical engineering from Graz University of Technology, Graz, Austria, in 2011, and the M.Sc. degree in sound and music computing from the Universitat Pompeu Fabra, Barcelona, Spain. She is currently pursuing the Ph.D. degree at the Department of Applied Mathematics, University of Seville, Seville, Spain. Her research interests include music information retrieval for computational ethnomusicology.

Emilia Gómez graduated as a Telecommunication Engineer specialized in signal processing from the Universidad de Sevilla, Sevilla, Spain. She received the DEA degree in acoustics, signal processing, and computer science applied to music (ATIAM) from IRCAM, Paris, and the Ph.D. degree in computer science and digital communication from the Universitat Pompeu Fabra (UPF), in July 2006. She is an Associate Professor (Serra-Húnter Fellow) with the Music Technology Group, UPF, Barcelona, Spain. Her research interests include melodic and tonal description of music audio signals, computer-assisted music analysis, and computational ethnomusicology.