PREDOMINANT FUNDAMENTAL FREQUENCY ESTIMATION VS SINGING VOICE SEPARATION FOR THE AUTOMATIC TRANSCRIPTION OF ACCOMPANIED FLAMENCO SINGING

E. Gómez[1], F. Cañadas[2], J. Salamon[1], J. Bonada[1], P. Vera[2] and P. Cabañas[2]
[1] Music Technology Group, Universitat Pompeu Fabra, Spain
[2] Telecommunication Engineering Department, University of Jaen, Spain
emilia.gomez@upf.edu, fcanadas@ujaen.es, justin.salamon@upf.edu, jordi.bonada@upf.edu, pvera@ujaen.es, pcabanas@ujaen.es

ABSTRACT

This work evaluates two strategies for predominant fundamental frequency (f_0) estimation in the context of melodic transcription from flamenco singing with guitar accompaniment. The first strategy extracts the f_0 from salient pitch contours computed from the mixed spectrum; the second separates the voice from the guitar and then performs monophonic f_0 estimation. We integrate both approaches with an automatic transcription system, which first estimates the tuning frequency and then implements an iterative strategy for note segmentation and labeling. We evaluate them on a flamenco music collection, including a wide range of singers and recording conditions. Both strategies achieve satisfying results. The separation-based approach yields a good overall accuracy (76.81%), although instrumental segments have to be manually located. The predominant f_0 estimator yields slightly higher accuracy (79.72%) but does not require any manual annotation. Furthermore, its accuracy increases (84.68%) if we adapt some algorithm parameters to each analyzed excerpt. Most transcription errors are due to incorrect f_0 estimations (typically octave and voicing errors in strong presence of guitar) and incorrect note segmentation in highly ornamented sections. Our study confirms the difficulty of transcribing flamenco singing and the need for repertoire-specific and assisted algorithms for improving state-of-the-art methods.

1. INTRODUCTION

Flamenco is a music tradition originating mostly from Andalusia in southern Spain. The singer has a main role and is often accompanied by the guitar and other instruments such as claps, rhythmic feet and percussion. This research aims to develop a method for computing detailed note transcriptions of flamenco singing from music recordings, which can then be processed for motive analysis or further simplified to obtain an overall melodic contour that will characterize the style. In this study we focus on accompanied singing, and propose a method comprised of two stages: predominant f_0 estimation and note segmentation. For the first stage, two alternative strategies are evaluated and compared: in the first, we use a state-of-the-art predominant f_0 estimation algorithm, which estimates the f_0 of the predominant melody directly from the full audio mix. In the second, we propose a source separation approach to isolate the singing voice and perform monophonic f_0 estimation.

2. SCIENTIFIC BACKGROUND

Automatic transcription is a key challenge in the music information retrieval (MIR) field. It consists of computing a symbolic musical representation from an audio recording. In polyphonic music material, there is an interest in transcribing the predominant melodic line [1]. Although we find some successful approaches for singing transcription [2,3], the singing voice is still one of the most complex instruments to transcribe, given the continuous character of the human voice and the variety of pitch ranges and timbre. Additional challenges in flamenco arise from the quality of recordings, the acoustic and expressive particularities of flamenco singing, its ornamental and improvisational character and the yet to be formalized musical structures [4].

In [5], we proposed a melodic transcription system from a cappella flamenco singing and we evaluated it against manual annotations of 72 performances. The obtained overall accuracy was around 70% (50 cents tolerance), which was significantly lower than the one obtained for a small test collection of pop/jazz excerpts ($\sim$85%). The study showed the importance of good monophonic f_0 estimation, and confirmed the difficulty of note segmentation for excerpts with unstable tuning or highly ornamented sections.

The goals of the present study are to apply this transcription system to accompanied singing and to perform a comparative evaluation of two alternative strategies for singing voice f_0 estimation. The first is to replace the monophonic f_0 detector by a predominant f_0 estimation method. The task of predominant f_0 estimation from polyphonic music (sometimes referred to simply as melody extraction) has received much attention from the research community in recent years, and state-of-the-art approaches yield an overall accuracy around 75% [6, 7]. A variety of different approaches have been proposed, based on tracking agents [8], classification [9], streaming rules [10] or pitch contour characterization [11]. The most common set of approaches are "salience based", i.e. they compute

a pitch salience representation from the audio signal, and then select the melody out of the peaks of this representation over time [8, 9, 11].

The second strategy is to separate the singing voice from the guitar accompaniment using source separation and transcribe the separated track. Recent singing voice separation methods can be classified into three categories: spectrogram factorization [12–14], pitch-based inference [15, 16] and repeating-structure removal [17]. Spectrogram factorization methods decompose a magnitude spectrogram as a set of components that represent features such as the spectral patterns (basis) or the activations (gains) of the active sources along the time. Fitzgerald and Gainza [12] propose a non-negative partial cofactorisation sharing a common set of frequency basis functions. In [13], an accompaniment model is designed, from the non-vocal segments, to fit the musical instruments and attempt separation of the vocals. In [14], the basis of the vocal track is learned from the mixture by keeping the accompaniment spectra fixed. Pitch-based inference methods use information from the pitch contour to determine the harmonic structures of singing voice. In [16], separation of both the voiced and the unvoiced singing voice is presented by means of the combination of detected unvoiced sounds and a spectral subtraction method to enhance voiced singing separation [15]. Repeating-structure removal methods [17] use a pattern recognition approach to identify and extract accompaniment segments, without manual labeling, which can be classified as repeating musical structures.

3. TRANSCRIPTION METHOD

Our method relies on two main stages: low-level feature extraction (mainly f_0) and note segmentation. We present the two alternatives for f_0 estimation compared in this study followed by a summary of the note segmentation approach.

3.1 Singing voice f_0 estimation

3.1.1 Predominant f_0 estimation

For predominant f_0 estimation, we use [11], which obtained the highest overall accuracy in MIREX 2011 [6]. First, the audio signal is analyzed and spectral peaks (sinusoids) are extracted. This process is comprised of three main steps: first a time-domain equal loudness filter is applied, which has been shown to attenuate spectral components belonging primarily to non-melody sources [19]. Next, the short-time Fourier transform is computed with a 46 ms Hann window, a hop size of 2.9 ms and a 4 zero padding factor. At each frame the local maxima (peaks) of the spectrum are detected. In the third step, the estimation of the spectral peaks' frequency and amplitude is refined by calculating each peak's instantaneous frequency (IF) using the phase vocoder method and re-estimating its amplitude based on the IF. The detected spectral peaks are subsequently used to compute a representation of pitch salience over time: a salience function. The salience function is based on harmonic summation with magnitude weighting, and spans a 5-octave range from 55Hz to 1760Hz. Details are provided in [11]. In the next stage, the peaks of the salience function are grouped over time using heuristics based on auditory streaming cues. This results in a set of pitch contours, out of which the contours belonging to the melody need to be selected. The contours are automatically analyzed and a set of contour characteristics is computed. In the final stage of the system, the contour characteristics and their distributions are used to filter out non-melody contours. The distribution of contour salience is used to filter out pitch contours at segments of the song where the melody is not present. Next, we obtain a rough estimate of the melodic pitch trajectory by computing a per-frame salience-weighted average of the remaining pitch contours and smoothing it over time using a sliding mean filter. This rough pitch trajectory is used to minimise octave errors (contours with the correct pitch class but in the wrong octave) and remove pitch outliers (contours representing highly unlikely jumps in the melody). Finally, the melody f_0 at each frame is selected out of the remaining pitch contours based on their salience. For further details the reader is referred to [11].

In addition to computing the melody f_0 sequence using the default algorithm parameters (denoted *MTG*), we also computed the melody adjusting three parameters of the algorithm for each musical excerpt: the minimum and maximum frequency threshold and the strictness of the voicing filter (cf. [11] for details). The results using the per-excerpt adjusted parameters are referred to as *MTGAdaptedparam*.

3.1.2 Singing voice separation and monophonic f_0 estimation

Standard Non-negative Matrix Factorization (NMF) [20] is not able to determine if a frequency basis belongs to a percussive, harmonic or vocal sound. Our proposal attempts to overcome this limitation without using any clustering process. A mixture spectrogram X is factorized into three separated spectrograms, X_p (percussive), X_h (harmonic) and X_v (vocal). Using similar spectro-temporal features [21, 22], harmonic sounds are modeled by sparseness in frequency and smoothness in time. Percussive sounds are modeled by smoothness in frequency and sparseness in time. Vocal sounds are modeled by sparseness in frequency and sparseness in time. Although it is not necessary to discriminate between percussive and harmonic sounds in the accompaniment, our experimental results showed we obtain better vocal separation using this discrimination. The proposed singing voice separation is composed of three stages: segmentation, training and separation.

In the segmentation stage, the mixture signal $X = X_{nonvocal} \bigcup X_{vocal}$ is manually labelled into vocal X_{vocal} (vocal+instruments) and non-vocal $X_{nonvocal}$ (only instruments) regions. In the training stage, from non-vocal regions, the percussive W_p and harmonic W_h basis vectors are learned using an unsupervised NMF percussive/harmonic separation approach based on spectro-temporal features.

$$X_{nonvocal} \approx X_p + X_h = W_p \cdot H_p + W_h \cdot H_h \qquad (1)$$

In the separation stage, the vocal spectrogram X_v is extracted from the vocal regions by keeping the percussive

W_p and harmonic W_h basis vectors fixed from the previous stage.

$$X_{vocal} \approx X'_p + X'_h + X_v = W_p \cdot H'_p + W_h \cdot H'_h + W_v \cdot H_v \quad (2)$$

In this manner, the singing voice signal $v(t)$ is synthesized from the vocal spectrogram X_v. To obtain an f_0 sequence from the synthesized voice signal, the traditional difference function is computed for each time frame index t:

$$d(\tau, t) = \sum_{n=0}^{W-1} (v(t+n) - v(t+n+\tau))^2 \quad (3)$$

where W is the length of the summation window and τ is the candidate pitch period. From this function, the cumulative mean normalized difference function can be computed as defined in [23]:

$$d_n(\tau, t) = \begin{cases} 1, & \tau = 0 \\ d(\tau, t)/[\frac{1}{\tau}\sum_{j=1}^{\tau} d(j, t)] & \text{otherwise.} \end{cases} \quad (4)$$

Observe that the function $d_n(\tau, t)$ can be viewed as a cost matrix, where each element (τ, t) indicates the cost of having a pitch period equal to τ at time frame t. We estimate the whole f_0 sequence by computing the lowest-cost path through the matrix $d_n(\tau, t)$. This computation is accomplished with dynamic programming. The endpoints of the path are fixed only for the t-axis and the path is constrained to advance step-by-step along t, under the condition $|\tau_{t-1} - \tau_t| \leq 1$. This condition ensures a continuous and smooth f_0 contour. The obtained f_0 is denoted as *UJA*.

3.2 Note segmentation and labeling

Our approach for note segmentation and labeling is adapted from a transcription system for mainstream popular music [18]. After consulting a group of flamenco experts from the COFLA project [1], we took the following design decisions. First, we define an equal-tempered scale with respect to an estimated tuning frequency. Second, we assume a constant tuning frequency value for each analyzed excerpt. Third, we transcribe all perceptible notes, including short ornamentations, in order to cover both expressive nuances and the overall melodic contour. We summarize below the mains steps of the transcription algorithm and we refer to [5] and [18] for further details.

3.2.1 Tuning frequency estimation

From the obtained f_0 envelope, we perform an estimation of the tuning frequency used by the singer assuming an equal-tempered scale. The tuning frequency is assumed to be constant for a given excerpt. We compute the maximum of the histogram of f_0 deviations from an equal-tempered scale tuned to 440 Hz. We then map the f_0 values of all frames into a single semitone interval with a one-cent resolution.

In our approach, we give more weight to frames where the included f_0 is stable by assigning higher weights to frames with low f_0 derivative. In order to smooth the resulting histogram and improve its robustness to noisy f_0 estimations, instead of adding a value to a single bin, we use a bell-shaped window that spans several bins. The maximum of this histogram (b_{max}) determines the tuning frequency deviation in cents from 440 Hz. The estimated tuning frequency in Hz then becomes

$$f_{ref} = 440 \cdot 2^{\frac{b_{max}}{1200}} \quad (5)$$

[1] http://mtg.upf.edu/research/projects/cofla

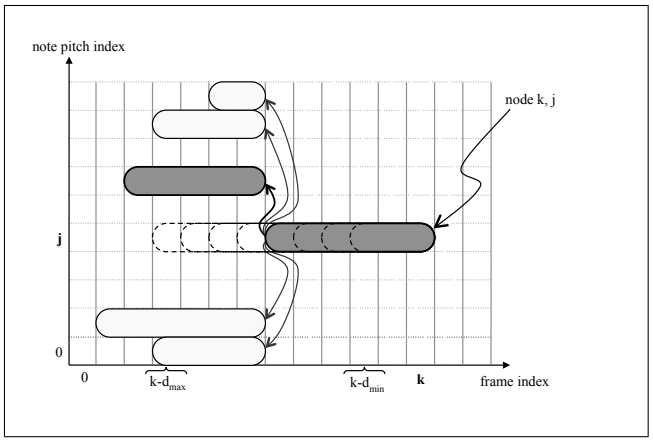

Figure 1. Matrix M used by the short note segmentation process, illustrating how the best path for a node with frame index k and note index j is determined. All possible note durations between d_{min} and d_{max} are considered, as well as all possible jumps to previous notes. The selected segmentation is marked with dark gray.

3.2.2 Short note transcription

The short note transcription step segments a single f_0 contour into notes. Using dynamic programming (DP), we find the note segmentation that maximizes a set of probability functions. The estimated segmentation corresponds to the optimal path among all possible paths along a 2-D matrix M (see Figure 1).

This matrix M has note pitches as rows and analysis frames as columns. Note pitches are quantized into semitones according to the estimated tuning frequency. Possible note pitches should cover the tessitura of the singer and include a $-\infty$ value for the unvoiced sections. Note durations are limited to a certain range $[d_{min}, d_{max}]$ of frames. The maximum duration d_{max} should be long enough so that it covers several periods of a vibrato with a low modulation frequency, e.g. $2.5Hz$, but also short enough to have good temporal resolution, e.g. avoid skipping short ornamentations.

Possible paths considered by the DP algorithm always start from the first frame, end at the last audio frame, and advance in time so that notes never overlap. A path p is defined by its sequence of N_p notes,
$p = \{n_{p0}, n_{p1}, \ldots, n_{pN_p-1}\}$, where each note n_{pi} begins at a certain frame k_{pi}, has a duration of d_{pi} frames and a pitch value of c_{pi}. The optimal path is defined as the path with maximum likelihood among all possible paths.

$$P = \arg\max_{p}\{L(p)\} \quad (6)$$

The likelihood $L(p)$ of a certain path p is determined as the product of likelihoods of each note $L(n_{pi})$ times the likelihood of each jump between consecutive notes $L(n_{pi-1}, n_{pi})$:

$$L(p) = L(n_{p0}) \cdot \prod_{i=1}^{N_{p-1}} L(n_{pi}) \cdot L(n_{pi-1}, n_{pi}) \quad (7)$$

In our approach, no particular characteristic is assumed a priori for the sung melody; therefore all possible note jumps have the same likelihood $L(n_{pi-1}, n_{pi}) = 1$. On the other hand, the likelihood of a note $L(n_{pi})$ is determined as the product of several likelihood functions based on the following criteria: duration (L_d), pitch (L_c), existence of voiced and unvoiced frames (L_v), and low-level features related to stability (L_s):

$$L(n_{pi}) = L_d(n_{pi}) \cdot L_c(n_{pi}) \cdot L_v(n_{pi}) \cdot L_s(n_{pi}) \quad (8)$$

Duration likelihood L_d is set so that it is small for short and long durations. Pitch likelihood L_c is defined so that it is higher the closer the frame f_0 values are to the note nominal pitch c_{pi}, giving more relevance to frames with low f_0 derivative values. The voicing likelihood L_v is defined so that segments with a high percentage of unvoiced frames are unlikely to be a voiced note, while segments with a high percentage of voiced frames are unlikely to be an unvoiced note. Finally, the stability likelihood L_s considers that a voiced note is unlikely to have fast and significant timbre or energy changes in the middle. Note that this is not in contradiction with smooth vowel changes, characteristic of flamenco singing.

3.2.3 Iterative note consolidation and tuning frequency refinement

The notes obtained in the previous step have a limited duration between $[d_{min}, d_{max}]$ frames, although longer notes are likely to have been sung. Therefore, it makes sense to consolidate consecutive voiced notes into longer notes if they have the same pitch. However, significant and fast energy or timbre changes around the note connection boundary may be indicative of phonetic changes unlikely to happen within a note, and thus may indicate that those consecutive notes are different ones. Thus, consecutive notes will be consolidated only if they have the same pitch and the stability measure of their connection L_s falls below a certain threshold.

Once notes are consolidated, it may be beneficial to use the note segmentation to refine the tuning frequency estimation. For this purpose, we compute a pitch deviation for each voiced note, and then estimate a new tuning frequency value from a one-semitone histogram of weighted note pitch deviations in similar way to that described in Section 3.2.1. The difference is that now we add a value for each voiced note instead of for each voiced frame. Weights are determined as a measure of the salience of each note, giving more weight to longer and louder notes. As a final step of this process, note nominal pitches are re-computed based on the new tuning frequency. This process is repeated until there are no more consolidations.

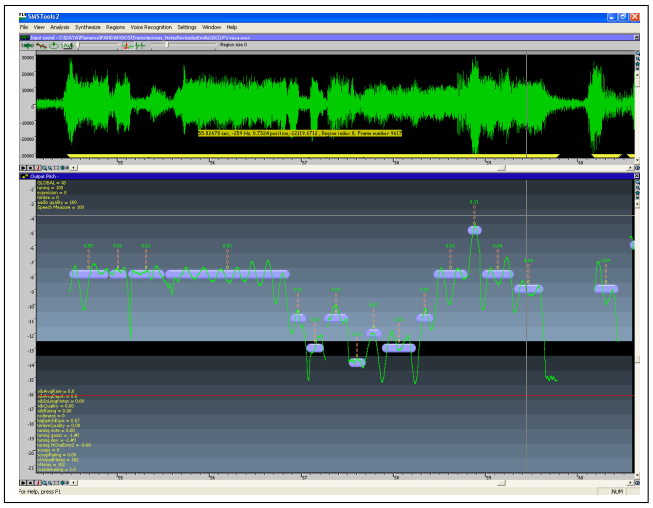

Figure 2. Visualization tool for melodic transcription. Audio waveform (top), estimated f_0 and pitch in a piano roll representation (bottom).

Figure 2 shows an example of a computed transcription. The system outputs both the extracted f_0 envelope and the estimated frame note pitch, according to an equal-tempered scale, as requested by flamenco experts for higher-level analyses.

4. EVALUATION STRATEGY

4.1 Music collection

We gathered 26.74 minutes of music, consisting of 30 performances of singing voice with guitar accompaniment (Fandango style). This collection has been built in the context of the COFLA project. It contains a variety of male and female singers and recording conditions. The average duration of the analyzed excerpts is 53.48 seconds and they contain a total of 271482 frames and 2392 notes.

4.2 Ground truth gathering

We collected manual note annotations from a musician with limited knowledge of flamenco music, so that there was no implicit knowledge applied in the transcription process. We provided him with the user interface shown in Figure 2. Since transcribing everything from scratch is very time consuming, we also provided the output of our transcription using the *MTGAdaptedParam* estimation as a guide. The annotator could listen to the original waveform and the synthesized transcription, while editing the melodic data until he was satisfied with the transcription. The criteria used to differentiate ornaments and pitch glides were discussed with two flamenco experts by collectively annotating a set of working examples, so that the annotator then followed a well-defined and consistent strategy.

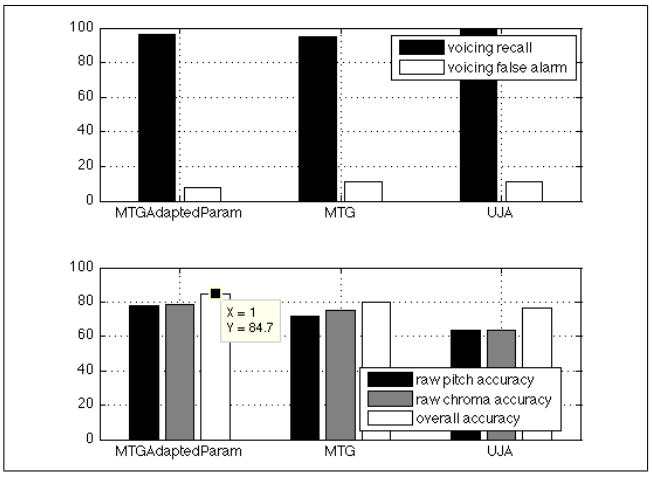

Figure 3. Frame-based accuracy measures (50 cents tolerance) for the considered approaches.

4.3 Evaluation measures

For evaluation we compute the measures used in the Audio Melody Extraction (AME) MIREX task [6]. The measures are based on a frame-by-frame comparison of the ground-truth to the estimated frequency sequence. Note that in our case we compare the ground truth to the frequency of the final note transcription, meaning any observed errors represent the combined errors introduced by the two stages of our method (f_0 estimation and note segmentation). Also, since we do not provide a pitch estimate for frames determined as unvoiced, incorrect voicing detection will also influence pitch accuracy (but not overall accuracy). We consider *Voicing recall*: % of voiced frames in the reference correctly estimated as voiced; *Voicing false alarm*: % of unvoiced frames in the reference mistakenly estimated as voiced; *Raw pitch accuracy*: % of voiced frames where the pitch estimate is correct within a certain threshold in cents (th); *Raw chroma accuracy*: same as the raw pitch accuracy except that octave errors are ignored; and *Overall accuracy*: total % of correctly estimated frames: correct pitch for voiced frames and correct labeling of unvoiced frames.

5. RESULTS

5.1 Frame-based pitch accuracy

Figure 3 shows the obtained accuracy measures for $th = 50$ cents. At first glance, we see that satisfying results are obtained for both strategies. The separation-based approach (*UJA*) yields good results (overall accuracy 76.81%, pitch accuracy 63.62%), as the guitar timbre can be accurately estimated from the instrumental segments. Nevertheless, these guitar segments have to be manually located. The predominant f_0 estimator (*MTG*) yields slightly higher overall accuracy (79.72%) and pitch accuracy (71.46%), and it does not require manual voicing annotation. Moreover, the overall accuracy increases to 84.68% (pitch accuracy 77.92%) if we adapt some algorithm parameters for each excerpt (*MTGAdaptedParam*). The observed voicing false alarm rate (around 10% for MTG and UJA) results from segments where the guitar is detected as melody.

Est.	Ref.	Vx Recall	Vx False Alarm	Raw pitch	Raw chroma	Overall accuracy
MTG	UJA	89.24	6.35	74.20	74.82	82.67
UJA	MTG	94.00	12.95	78.29	78.93	82.65

Table 1. Accuracy measures between f_0 estimations.

The obtained results are slightly higher than the ones obtained for a cappella singing [5] when considering the same note segmentation algorithm together with a monophonic f_0 estimator. This is due to two main reasons. Primarily, as the singer follows the tuning reference of the guitar, there are no tuning errors and the note labeling results are improved. Also, as the voice is very predominant with respect to the guitar, the predominant f_0 estimation method works very well for this material.

5.2 Agreement between f_0 estimations

We also estimate the agreement between both f_0 strategies by computing the evaluation measures with one estimator as ground truth and the other one as estimation. Results are presented in Table 1. We observe that in both cases the overall agreement is around 82.5%. The main difference between the approaches is in the determination of voiced sections. Whilst in *UJA* only large non-voiced sections were manually annotated, *MTGAdaptedParam* also attempts to automatically detect shorter unvoiced sections in the middle of the piece.

5.3 Error analysis

We observe that for the two considered strategies, transcription errors are introduced in both stages of the transcription process (f_0 estimation and note segmentation).

Regarding singing voice f_0 estimation, voicing seems to be the main aspect to improve. Voicing false positives occasionally appear during melodic guitar segments and in short unvoiced phonemes (e.g. fricatives). On the other hand, the singing voice f_0 is sometimes missed in the presence of strong instrumental accompaniment, resulting in voicing false negatives. Since the subsequent note segmentation stage relies on the voicing estimation, voicing errors during the f_0 estimation are bound to introduce errors in the note segmentation stage as well. Another type of error is fifth or octave errors at segments with highly predominant accompaniment. This occurs especially with the *UJA* method, as low harmonics of the singing voice might be erased from the spectrum during the separation process.

Regarding the note segmentation algorithm, most of the errors happen for short notes; either they are consolidated while the annotation consists of several close notes, or vice versa. This especially happens where the energy envelope also accounts for the presence of guitar, so that onset estimation becomes more difficult. Finally, some of the errors occur due to wrong pitch labeling of very short notes, as the f_0 contour is short and unstable. This demonstrates the difficulty of obtaining accurate note transcriptions for flamenco singing, given its ornamental character and the

continuous variations of f_0, easily confused with deep vibrato or pitch glides. The great variability of the vocal f_0 contour can be observed in Figure 2.

6. CONCLUSIONS

This paper presents an approach for computer-assisted transcription of accompanied flamenco singing. It is based on an iterative note segmentation and labelling technique from f_0, energy and timbre. Two different strategies for singing voice f_0 estimation were evaluated on 30 minutes of flamenco music, obtaining promising results which are comparable to (and even better than) previous results for monophonic singing transcription. The main sources of transcription errors were identified: in the first stage (f_0 estimation) the main issue is voicing detection (e.g. identification of the guitar as voice), though we occasionally observe pitch errors (e.g. wrong f_0 in the presence of guitar) as well. In the second stage (note segmentation) we observed errors in segmenting short notes and labeling notes with an unstable f_0 contour. There is still much room for improvement. One limitation of this work is the small amount of manual annotations. This is due to the fact that manual annotation is very time consuming and difficult to obtain, and has a degree of subjectivity. We are currently expanding the amount of manual annotations. The second limitation is that we only have manual annotations on a note level (quantized to 12 semitones) and not the continuous f_0 ground truth, which would allow us to evaluate separately the accuracy of the two main stages of the algorithm. We plan to work on this issue. Finally, we plan to quantify the uncertainty of the ground truth information by comparing annotations in different contexts, and adapt the algorithm parameters accordingly.

7. ACKNOWLEDGEMENTS

The authors would like to thank the COFLA[1] team for providing the data set and expert knowledge in flamenco music. This work has been partially funded by AGAUR (mobility grant), the COFLA project (P09-TIC-4840 Proyecto de Excelencia, Junta de Andalucía) and the Programa de Formación del Profesorado Universitario of the Ministerio de Educación de España.

8. REFERENCES

[1] A. Klapuri, and M. Davy (Eds): "Signal Processing Methods for Music Transcription," Springer-Verlag, 2006.

[2] T. Mulder, J. P. Martens, M. Lesaffre, M. Leman, B. De Baets, and H. De Meyer: "An Auditory Model Based Transcriber of Vocal Queries," Proc. of ISMIR, 2003.

[3] M. P. Ryynänen: "Singing transcription," in Signal processing methods for music transcription (A. Klapuri and M. Davy, eds.), Springer, 2006.

[4] J. Mora, F. Gomez, E. Gómez, F. Escobar-Borrego, and J.M. Diaz-Bañez: "Characterization and melodic similarity of a Cappella flamenco cantes," Proc. of ISMIR 2010.

[5] E. Gómez, J. Bonada, and J. Salamon: "Automatic Transcription of Flamenco Singing from Monophonic and Polyphonic Music Recordings," Proc. of FMA, 2012.

[6] J. Salamon, E. Gómez: "Melody Extraction from Polyphonic Music: MIREX 2011," in Music Information Retrieval Evaluation eXchange (MIREX), 2011.

[7] G. E. Poliner, D. P. W. Ellis, F. Ehmann, E. Gómez, S. Streich, and B. Ong: "Melody transcription from music audio: Approaches and evaluation," IEEE Transactions on Audio, Speech and Language Processing, 15(4):1247:1256, 2007.

[8] M. Goto: "A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals," Speech Communication, 43:311:329, 2004.

[9] G. E. Poliner and D.P.W. Ellis: "A Classification Approach to Melody Transcription," Proc. of ISMIR, 2005.

[10] K. Dressler: "An auditory streaming approach for melody extraction from polyphonic music," Proc. Of ISMIR, pp. 19:24, Miami, 2011.

[11] J. Salamon, and E. Gómez: "Melody Extraction from Polyphonic Music Signals using Pitch Contours Characteristics," IEEE Transactions on Audio, Speech and Language Processing, 20(6):1759-1770, August 2012.

[12] D. FitzGerald, and M. Gainza: "Single Channel Vocal Separation using Median Filtering and Factorisation Techniques," ISAST Transactions on Electronic and Signal Processing, 4(1):62-73, 2010 (ISSN 1797-2329)

[13] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval: "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," IEEE Transactions on Audio, Speech, and Language Processing, 15(5): 1564:1578, July 2007.

[14] B. Raj, P. Smaragdis, M. V. Shashanka, and R. Singh: "Separating a foreground singer from background music," Proc. Int Symp. Frontiers Res. Speech Music (FRSM), India, 2007.

[15] Y. Li, and D. Wang: "Separation of singing voice from music accompaniment for monaural recordings," Proc. of ICASSP, 15(4):1475:1487, May, 2007.

[16] H. Chao-Ling, and R. Jyh-Shing: "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," IEEE Transactions on Audio, Speech, and Language Processing, 18(2):310:319, February 2010.

[17] Z. Rafii, and B. Pardo: "A Simple Music/Voice Separation Method based on the Extraction of the Repeating Musical Structure," Proc. of ICASSP, Prague, May, 2011.

[18] J. Janer, J. Bonada, M. de Boer, and A. Loscos: "Audio Recording Analysis and Rating," Patent pending US20080026977, Universitat Pompeu Fabra, 06/02/2008.

[19] J. Salamon, E. Gómez, and J. Bonada: "Sinusoid Extraction and Salience Function Design for Predominant Melody Estimation," Proc. of DAFX, Paris, 2011, pp. 73-80.

[20] D. Lee, and H. Seung: "Algorithms for Non-negative Matrix Factorization," Advances in NIPS, pp. 556-562, 2000

[21] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama: "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," Proc. of EUSIPCO, 2008

[22] T. Virtanen: "Monaural Sound Source Separation by Non-Negative Matrix Factorization with Temporal Continuity and Sparseness Criteria," IEEE Transactions on Audio, Speech, and Language Processing, 3(15), March 2007.

[23] A. de Cheveigné, and H. Kawahara: "YIN, a Fundamental Frequency Estimator for Speech and Music," Journal of the Acoustic Society of America, 111(4):1971:1930, 2002.