Towards Computer-Assisted Flamenco Transcription: An Experimental Comparison of Automatic Transcription Algorithms As Applied to A Cappella Singing

Emilia Gómez and Jordi Bonada

Music Technology Group, Department of Information and Communication Technologies, Universitat Pompeu Fabra

Roc Boronat, 138

08018 Barcelona, Spain

{emilia.gomez,jordi.bonada}@upf.edu

Abstract

This paper deals with automatic transcription of flamenco music recordings, more specifically a cappella singing. We first study the specificities of flamenco singing and propose a transcription system based on fundamental frequency and energy estimation, which incorporates an iterative strategy for note segmentation and labelling. The proposed approach is evaluated on a music collection of 72 performances, including a variety of singers, recording conditions, presence of percussion, background voices and noise. We obtain satisfying results for the different tested approaches and our system outperforms a state-of-the-art approach designed for other singing styles. In this study, we discuss the difficulties found in transcribing flamenco singing and in evaluating the obtained transcriptions, we analyze the influence of the different steps of the algorithm, and we state the main limitations of our approach and discuss the challenges for future studies.

Introduction

This section presents the motivation, scientific background and the main goals of this study.

Flamenco and its musical transcription

Flamenco is a music tradition originally from Andalusia in southern Spain. We refer to the books of Blas Vega and Ríos Ruiz (1988), Navarro and Ropero (1995), and Gamboa (2005) for a comprehensive study of styles, musical forms and history of flamenco. Flamenco music germinated and nourished mainly from the singing tradition (Gamboa 2005). Accordingly, the singer's role soon became dominant and fundamental. In the flamenco jargon, singing is called *cante*, and songs are termed *cantes*. The flamenco guitar can also accompany the singer; other instruments include claps, rhythmic feet and percussion. The origin and evolution of the different flamenco styles (*palos*) and variations have been studied by different disciplines, including ethnomusicology, literature and anthropology (Katz 2006). Each flamenco style is characterized by a certain melodic skeleton, which can be subject to ornamentation and variation (Donnier 1996).

Because of its oral transmission, there are no written flamenco scores. Flamenco experts have put much effort into generating manual transcriptions after listening to live performances or field recordings, as a means to catalogue, classify and imitate the most relevant performers and their stylistic traits (Hurtado and Hurtado 1998); (Hurtado and Hurtado 2002); (Fernández 2004); (Hoces 2011). As pointed out by Toiviainen and Eerola (2006) and Lesaffre et al. (2004) in other contexts, manual analyses provide very accurate and expert information, but they are very time consuming and subjective or prone to inconsistencies. This is also the case in flamenco, due to two main reasons. First, there is a disagreement on the most adequate transcription methodology; second, there is a degree of subjectivity in the transcription process.

Regarding the transcription format, Hurtado and Hurtado (1998, 2002) forcefully argue for the use of Western notation. On the contrary, Donnier (1996) suggests the adaptation of plainchant neumes. In his work, Donnier proposes a methodology for flamenco transcription based on four different levels,

depending on the target usage (Donnier 2011). The first level corresponds to a spectral representation (time vs frequency) and provides detailed information about the instantaneous evolution of pitch, energy and timbre. The second level converts it into a non-continuous sequence of notes (pitches are represented in a score but he discards duration information), considering only the main notes of the melody and aligning the analyzed recording with other performances of the same style. According to Donnier (2011), one or several flamenco experts should jointly carry out this second level of description. The third level converts the previous representation to neumatic notation based on (Cardine 1970) as illustrated in Figure 1, and it also requires expert knowledge. The fourth level finally implies further simplification of the transcription in order to extract the style's melodic skeleton.

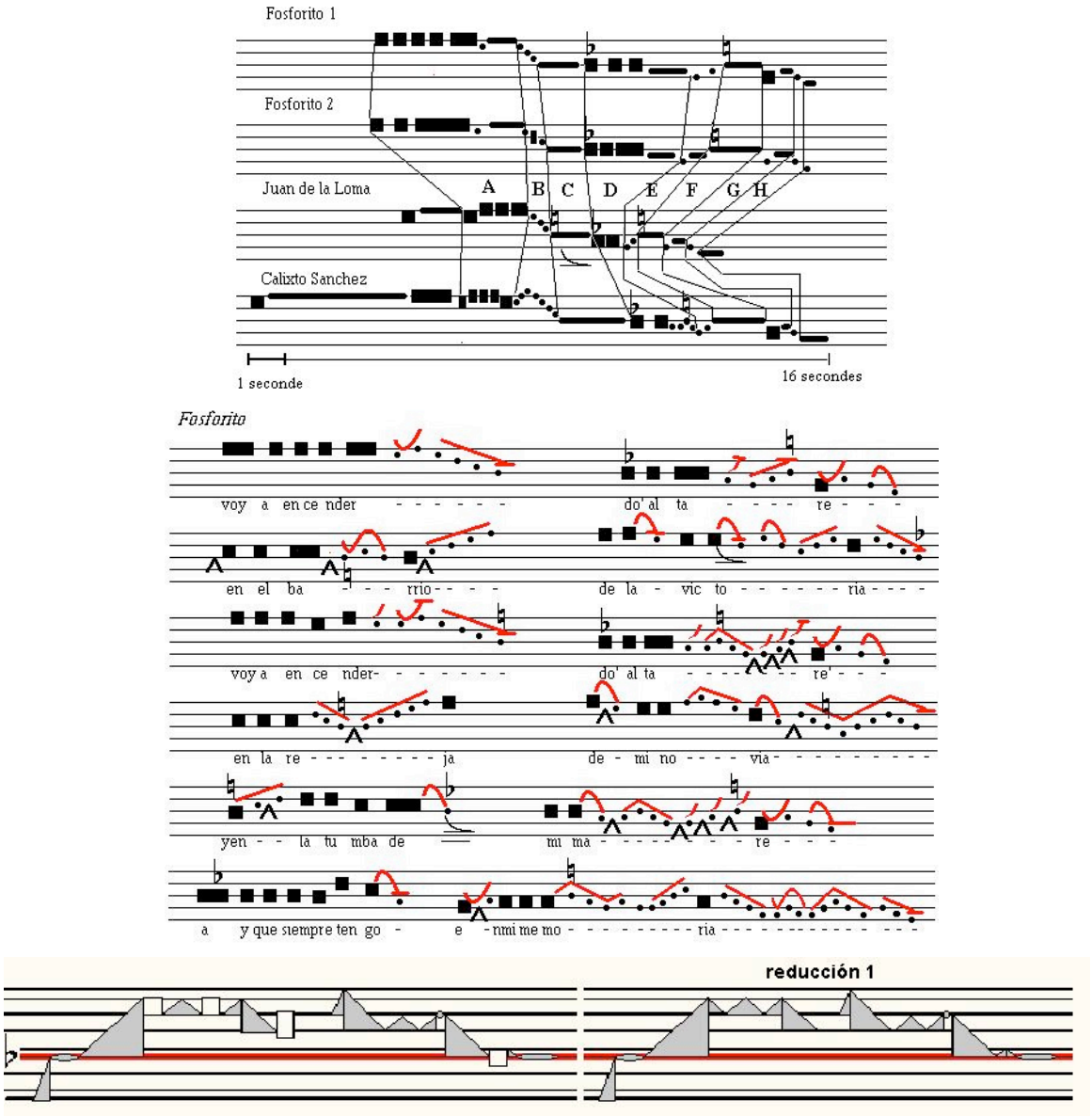

Figure 1: Example of manual transcription by Dr. Philippe Donnier. Second (top), third (middle) and fourth level (bottom). From Donnier (2011), reproduced with the author's permission.

This proposal for transcription methodology illustrates the fact that, even if experts agree on the use of a certain format, there is a degree of subjectivity in the transcription process, partially due to the high degree of ornamentation in flamenco music and the lack of common criteria.

Audio content analysis technologies could provide means for generating flamenco transcriptions. First, they would allow the computation of melodic features in different abstraction levels and help to set up a standard methodology for transcription. Second, they could facilitate the analysis of large audio collections.

Automatic transcription of sung melodies

Automatic transcription is one of the main research challenges in the field of sound and music computing. It consists of computing a symbolic musical representation (in terms of Western notation) from a given musical performance (Klapuri 2006). For monophonic music material, the obtained transcription is a single musical line, usually a melody (Gómez et al. 2003) and in polyphonic music material there is an interest in transcribing the predominant melodic line (Klapuri 2006). Transcription systems can provide melodic descriptors at different levels. Low-level features mostly related to melody are energy, associated with loudness, and fundamental frequency (f0) related to its perceptual correlate, pitch. From now on, we will use the term pitch referring to f0. In a higher structural level, audio streams are segmented into notes, and their duration and pitch provide a symbolic representation. This representation can be the input to higher-level music analyses, e.g. ornament detection, melodic contour extraction or key or scale analysis. Current systems for automatic transcription are usually structured as three different stages, represented in Figure 2: low-level (frame-based) descriptor extraction, note segmentation and note labelling.

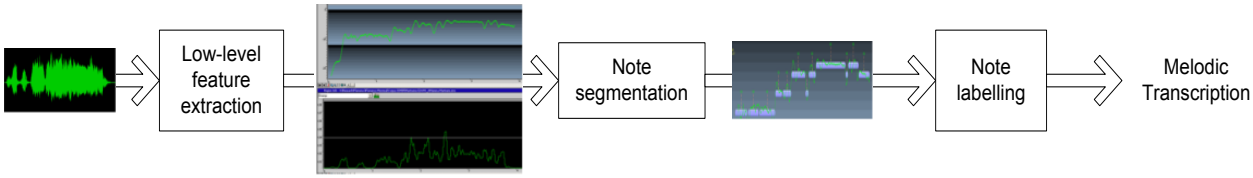

Figure 2: Steps for automatic transcription

When dealing with monophonic music signals, existing systems provide satisfying results for a great number of musical instruments. Although we find some successful approaches for singing voice (Mulder et al. 2003; Ryynänen 2006), it is still one of the most complex instruments to transcribe, even in a monophonic context. This is due to several factors, such as the continuous character of the human voice and the variety of pitch ranges and timbre. Many difficulties then appear for obtaining correct f0 estimations, detecting note transitions (onsets and offsets) and labelling notes in terms of pitch or duration. These difficulties are verified when comparing state-of-the-art systems for audio onset detection (task related to note segmentation and required for automatic transcription), which yield an average F-measure (a statistical measure of accuracy, from 0 to 1) around 0.78 according to the 2010 edition of the Music Information Retrieval Evaluation eXchange (MIREX). This F-measure is obtained for a mixed dataset of 85 files, but if we just consider the 5 tested singing voice excerpts, the maximum F-measure is 0.47. This suggests that state-of-the-art systems for singing voice transcription are not accurate enough to be used in an unsupervised way.

In addition, current approaches are oriented towards mainstream popular music. We might then ask how they would perform for, e.g. traditional music. Existing literature has addressed the problem of music transcription in repertoires from different music traditions, e.g. (Six and Cornelis 2011; Gedik et al. 2012; Ness et al. 2010; Widwans and Rao 2012). We focus here on flamenco singing, which presents some particularities that further complicate its automatic transcription, as we discuss below.

Goals and structure of the paper

This paper addresses two main challenges in automatic flamenco transcription. The first one is the limitation of current technologies to deal with traditional music in general and flamenco singing in particular. Second, the fact

that state of the art tools are not reliable enough when contrasted with expert analyses.

Our final goal is to build a tool for 'computer-assisted' transcription, where state of the art audio analysis technologies assist flamenco experts in the transcription of flamenco a cappella singing by providing accurate f0 estimation and note sequences. In this study, we evaluate different approaches for f0 estimation and incorporate an iterative strategy for note segmentation and labelling. The proposed system is used to generate, in a semi-automatic way, a corpus of transcriptions from a cappella singing styles. As a result of this process, we evaluate how current algorithms work in this context and we discuss the difficulty of evaluation and the technological limitations that should be addressed in future research.

The paper is structured as follows. We first review the specificities of flamenco singing and then present our approach for automatic transcription. We evaluate its performance by gathering manual transcriptions, and we analyze the influence of the different algorithmic steps. Finally we study the main limitations of our approach and discuss future improvements that would be required.

Characteristics of flamenco singing

Flamenco singing presents several characteristics that differentiate it from other styles. Several of these traits are related to the piece melodic structure, others relate to the expressive resources of a particular singer and others to their voice quality and timbre. The unavailability of scores and the oral character of flamenco have made discrimination between these diverse types of expressive assets very difficult. This is linked to a long-standing discussion in the flamenco community around the definition of *styles* and *variants*.

We introduce here some of the most relevant features of flamenco sung melodies (Fernández 2004; Mora et al. 2010). Regarding melody, flamenco singing is characterized by the use of a small pitch range or tessitura, usually

limited to one octave. Melodies are characterized by the insistence on a particular pitch and its neighbors. In addition, flamenco presents a high degree of ornamentation (*melisma),* produced by continuous variations of pitch and/or energy, and easily confused with expressive resources such as deep vibrato, portamenti or pitch glides.

If we focus on timbre we find that, depending on the period, there are some trends in timbre characteristics of flamenco singers. In general, there is a distinction between *Gypsy* and *non-Gypsy* voices, but we cannot establish a unique or characteristic flamenco timbre. Voices generally present a strong breathiness, i.e. presence of air, and there is an absence of a high frequency (singer) formant, which is characteristic of classical singing.

Moreover, the evolution of pitch, energy and timbre in flamenco singing is also characterized by its instability, which is often used as an expressive resource. For instance, tuning is not meant to be accurate for flamenco singers, and it often varies along time in a cappella singing (as singers do not have a reference). In addition, some notes can be mistuned. Finally, the singers' note attacks are not always clear; there are often smooth transitions between notes.

The aforementioned characteristics, some of which are illustrated in Figure 3, are quite in contrast with classical singing styles, where it is very important to achieve good tuning and timing, and where the timbre is characterized by its stability, absence of air, and the presence of high-frequency formants (i.e. the singer formant) (Sundberg 1987).

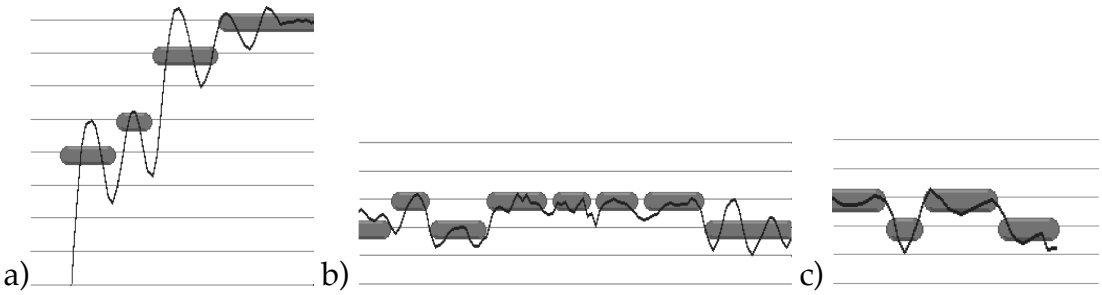

Figure 3: Examples of characteristic pitch envelopes in a cappella

flamenco singing (a. note beginning. b. middle ornamentations. c. note ending). The ovals behind the pitch indicate the transcriptions obtained by the proposed system.

We finally mention two additional difficulties found in the analysis of flamenco a cappella singing. First, existing historical recordings are of bad quality and some of them contain a high amount of reverberation and noise. Second, although the music is assumed to be monophonic, many recordings incorporate some percussion instruments, clapping and accompanying voices.

Selected approach

Figure 4 shows an overall diagram of our approach for automatic transcription. In the following sections we describe the different steps of the algorithm.

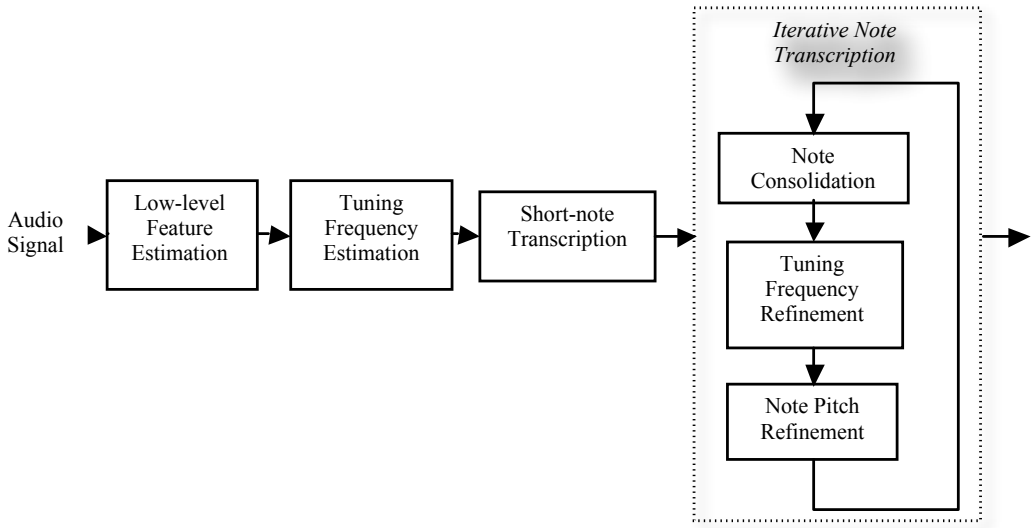

Figure 4: Diagram for automatic transcription.

Melodic representation

After consulting a group of flamenco experts, we defined three levels of melodic representation: (1) performance level including expressive resources (e.g. vibrato, portamenti) captured by instantaneous pitch and energy envelops; (2) detailed transcription, including ornaments (sequence of short perceptible notes); and (3) melodic contour, capturing the main notes of the melody (often long and

stable). In the context of this research, the third level of representation is generated by flamenco experts, as suggested by Donnier (2011), based on the automatic extraction of the first and second levels of description. The explanation of this manual transcription procedure is out of the scope of this paper.

The flamenco experts considered that expressive inflexions (1) were not relevant for the final transcription and style characterization, although they were valuable for singer style characterization and intonation analysis. Moreover, they proposed to use an equal-tempered scale for note transcription, i.e. note pitches are quantized to an equal-tempered chromatic scale with respect to an estimated tuning frequency (e.g. 440 Hz). As we analyze musical phrases, we assume a constant tuning frequency value for each analyzed excerpt. That means that even if the singer sings out of tune or does not maintain a constant tuning, the system discards this mistuning for transcription, i.e. it quantizes each note's f0 to the closest semitone with respect to the estimated tuning frequency.

The system then outputs low-level frame-based features (energy and fundamental frequency), which can be used for expressive characterization (1) and detailed transcriptions (2) including note descriptors for all perceptible notes (energy, onset position, duration and pitch). It also outputs a global descriptor (tuning frequency).

Low-level feature estimation

The audio signal is first divided into overlapping frames 50 msec long, with ~5.8 msec between frame onsets (~172 frames/second). From each frame, we compute its spectrum, energy and estimate its fundamental frequency (f0). Many algorithms have been proposed in the literature for estimating the f0 of monophonic music signals, both in the time and frequency domain and adapted to different instruments. We refer to (Gómez et alt. 2003) or (Klapuri 2006) for a detailed review. In order to analyze the influence of the chosen f0 estimation algorithm on the final behaviour of the system, we have integrated in our system

three state-of-the-art approaches: 1. Time-domain autocorrelation: we have considered the well-known yin algorithm proposed by de Cheveigné and Kawahara (2002) (*yin*); 2. Frequency-domain harmonic matching: we have implemented an algorithm based on the Two-Way Mismatch algorithm (*twm*) proposed by Maher and Beauchamp (1994), as presented in (Cano 1998). This algorithm tries to match the spectral peaks (local maxima of the spectrum) to a harmonic series; 3. Frequency-domain autocorrelation: the third method we consider is our own *sac* (Spectrum AutoCorrelation), which is based on the computation of amplitude correlation in the frequency domain. Regarding the choice of parameters we did not empirically investigate an optimal choice for our particular data set. For the *twm* and *sac* algorithms, we set the allowed f0 range was set to [60,1100] Hz. We did not modify the default parameters of the *yin* algorithm in an implementation by its original author, where f0 range was set to [30,fs/4], the window length to ~33 ms, hop size to ~0.72 ms, and the aperiodicity threshold to 0.1. We resampled the output f0 vector to match the frame rate used by our system (~172 frames/second). Since the *sac* method has not been published previously, we will now describe its main characteristics.

SAC algorithm

This algorithm consists of three steps: resampling, frame-by-frame f0 candidate estimation, and post-processing. First the audio signal is downsampled using polyphase filters to a sampling frequency around 11Khz, in order to reduce the computational cost. Next the audio is segmented into a sequence of overlapping frames of ~50 ms with a rate of ~172 frames per second. Then we take each frame and compute its spectrum with a Blackman-Harris window without zero-padding. We then convolve the complex spectrum with a triangular kernel of variable length: from 0 Hz length at 0 Hz to 72 Hz length at 1 kHz and above. This filtering process generates a multi-resolution spectrum with desirable properties. Next we compute the amplitude of the obtained spectrum in dB scale

and estimate a smooth average by convolving it with a triangular window of variable length: from 80 Hz length at 0 Hz to 180 Hz length at 700 Hz and above. Afterwards we compute the difference function (noted as Y) between the amplitude spectrum and its smooth version. Finally we compute the autocorrelation function of $D=[max(Y,0)]^2$ and normalize it by its maximum value.

We estimate the local maxima of the normalized autocorrelation function and choose as the f0 candidate the one with minimum frequency that fulfills several heuristic rules (e.g. have a significant autocorrelation value compared to the maximum value, be within the allowed f0 range). Finally, we refine the f0 candidate frequency using a 2nd order polynomial.

The last step consists of a post-processing that smoothes the estimated f0 function with a short latency of a few frames, corrects octave jumps, and decides if a frame is voiced or not. In order to help the voiced/unvoiced decision we compute a voiciness probability combining several heuristic rules (e.g. a voiced frame is not likely to have very low energy, nor a high number of time-domain zero crossings, nor very low autocorrelation values).

Manual annotations of f0 envelope

In order to measure the amount of errors caused by wrong f0 estimation in the final performance, we have also introduced a manually edited f0 envelope (*Corrected-f0*). This envelope was obtained by manual edition of the last approach (*sac*), where we manually corrected the most relevant f0 errors, mainly caused by reverberation (end of phrases) and noise (background voices and percussion).

Note segmentation and labelling

The algorithm for note segmentation and labelling is based on the one described in (Janer et al. 2008), and consists of three main steps: tuning frequency estimation, transcription into short notes, and an iterative process involving note consolidation and refinement of the tuning frequency.

Tuning frequency estimation

As we analyze singing voice performances, the frequency reference used by the singer to tune the piece is unknown. In order to locate the main pitches, we perform an initial estimation of this tuning frequency assuming an equal-tempered scale. We also assume that this reference frequency is constant for the analyzed excerpt. We estimate it by computing the maximum of the histogram of f0 deviations from an equal-tempered scale tuned to 440 Hz. This histogram represents the mapping of f0 values of all frames into a single semitone interval with a resolution of c_{res}=1 cent. The value added to the histogram is a weight representing the relevance of each frame. In our approach, we give more weight to frames where the included f0 is stable by assigning higher weights to frames where the values of the f0 derivative are low. In order to smooth the resulting histogram and improve its robustness to noisy f0 estimations, instead of adding a value to a single bin, we use a bell-shaped window that spans 20 cents. Note that since the histogram axis is wrapped onto a one-semitone deviation, adding a window around a boundary of the histogram would contribute also to the other boundary. The maximum of this histogram (c_{ref}) determines the tuning frequency deviation in cents from 440 Hz. Therefore, the estimated tuning frequency in Hz becomes $f_{ref} = 440 \cdot 2^{\frac{c_{ref}}{1200}}$.

Short note transcription:

As a second step, the audio signal is segmented into short notes by finding the segmentation that maximizes a certain likelihood function. The estimated segmentation corresponds to the optimal path among all possible paths along a 2-D matrix M (see Figure 4). This matrix has the possible note pitches in cents as rows ($[c_0,c_n]$, with one semitone resolution) and the analysis frame times as columns. Note that the possible note pitches should cover the tessitura of the singer ($[c_{min},c_{max}]$) and include a $-\infty$ value for the unvoiced sections. In this step, note durations are limited to a certain range between n_{min} and n_{max} frames. The

maximum duration n_{max} should be long enough to cover several periods of a vibrato with a low modulation frequency, e.g. 2.5 Hz, but short enough to have a good temporal resolution, for example, a resolution that avoids skipping fast notes with a very short duration. In a later step, successive short notes having the same pitch can be consolidated into longer ones.

Possible paths considered by the 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 m notes, $P=\{N_0,N_1,\ldots,N_{m-1}\}$, where each note N_i begins at a certain frame k_i, has a pitch deviation of c_i in cents relative to the tuning reference c_{ref}, and a duration of n_i frames. The optimal path is defined as the path with maximum likelihood among all possible paths. The likelihood L_P of a certain path is determined as the product of likelihoods of each note L_{N_i} by the likelihood of each jump (i.e. connection) between consecutive notes L_{N_i-1,N_i}, that is

$$L_P = L_{N_0} \cdot \prod_{i=1}^{m-1} L_{N_i} \cdot L_{N_i-1,N_i} .$$

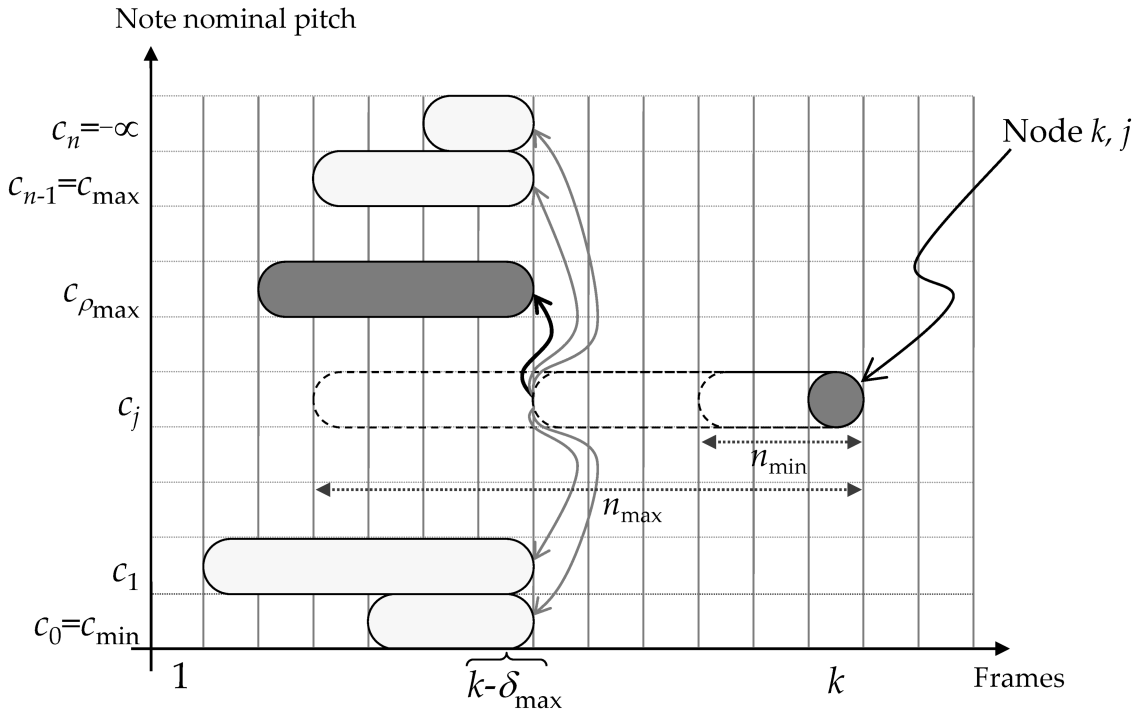

Figure 5: This figure shows the matrix M used by the short note segmentation process, and illustrates how the best path is determined for the node with frame k and note j (corresponding to pitch c_j). All possible note durations between n_{min} and n_{max} are considered, as well as all possible jumps (i.e. connections) to previous notes (as shown by the curved gray or black arrows in the middle of the figure). In this example δ_{max} is found to be the most likely note duration and ρ_{max} the pitch index of the previous note (as shown by the black curved arrow in the middle of the figure).

In order to speed up the process, an approximate optimal path can be found by advancing through the matrix columns from left to right. For each k^{th} column (frames), we decide at each j^{th} row (note pitches) an optimal note duration and jump by maximizing the note likelihood times the jump likelihood times the previous note's accumulated likelihood, among all combinations of possible note durations and jumps. This maximized likelihood is then stored as the accumulated likelihood $\hat{L}_{k,j}$ for that node of the matrix, together with the

optimal note duration and jump. Therefore,

$$\hat{L}_{k,j} = L_{N_{k,j}}\left(\delta_{\max}\right)\cdot L_{N_{k-\delta_{\max},\rho_{\max}},N_{k,j}}\cdot\hat{L}_{k-\delta_{\max},\rho_{\max}}$$
$$= \max\left(L_{N_{k,j}}\left(\delta\right)\cdot L_{N_{k-\delta,\rho},N_{k,j}}\cdot\hat{L}_{k-\delta,\rho}\right) \quad ,$$
$$, \quad \forall\delta\in\left[n_{\min},n_{\max}\right], \forall\rho\in\left[0,n\right]$$

where δ represents the note duration in frames, and ρ the row index of the previous note using zero-based indexing. For the first column, the accumulated likelihood is 1 for all rows ($\hat{L}_{0,j}=1, \forall j\left[0,n\right]$). The optimal path of the matrix $P_{\max}$ is obtained by first finding the node of the last column with a maximum accumulated likelihood, and then by following its corresponding jump and note sequence.

In our approach, no particular characteristic is assumed *a priori* for the sung melody; therefore all note jumps have the same likelihood $L_{N_{i-1},N_i}=1$, $\forall i\in\left[1,m-1\right]$. On the other hand, the likelihood L_{N_i} of a note N_i is determined as the product of several likelihood functions based on the following criteria: duration (L_{dur}), fundamental frequency (L_{pitch}), existence of voiced and unvoiced frames (L_{voicing}), and low-level features related to stability ($L_{\text{stability}}$). For the note N_i, its likelihood L_{N_i} is computed as $L_{N_i}=L_{\text{dur}}\cdot L_{\text{pitch}}\cdot L_{\text{voicing}}\cdot L_{\text{stability}}$. The duration likelihood L_{dur} is set so that it is small for very short and very long durations. The pitch likelihood L_{pitch} is determined so that the likelihood is higher the closer the estimated pitch contour values are to the note nominal pitch c_i and vice versa, giving more relevance to frames with lower values for the first derivative of the pitch contour. The voicing likelihood L_{voicing} is determined 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 considers that a voiced note is unlikely to have fast and significant timbre or energy changes in the middle of the note. Note that this is not in contradiction with the typical characteristic of flamenco singing of changing the vowel at note endings, since those changes are

mostly smooth.

Iterative note consolidation and tuning frequency refinement:

In this third step, consecutive notes with the same pitch and a smooth transition are consolidated, the estimated tuning frequency is refined according to the obtained notes, and the note nominal pitch is re-estimated according to the new tuning frequency. This whole process is repeated until there are no more consolidations.

Note consolidation: the "short notes" obtained in the previous step have a limited duration between $n_{\min}$ and $n_{\max}$, 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 falls below a certain threshold.

Tuning frequency refinement: In the first step, tuning frequency was estimated from the fundamental frequency contour. However, once notes have been segmented, it may be beneficial to use the note segmentation to refine the tuning frequency. For this purpose, we compute a pitch deviation for each voiced note, and then estimate the new tuning frequency from a one-semitone histogram of weighted note pitch deviations similar to the first step. 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.

Figure 6 shows an example of the whole system output. As mentioned above, the system transcribes according to an equal-tempered scale, as requested by flamenco experts. This means that, even if the performer is out of tune, we

approximate the used pitches to an equal-tempered chromatic scale according to the estimating tuning frequency, i.e., mistuning is not transcribed.

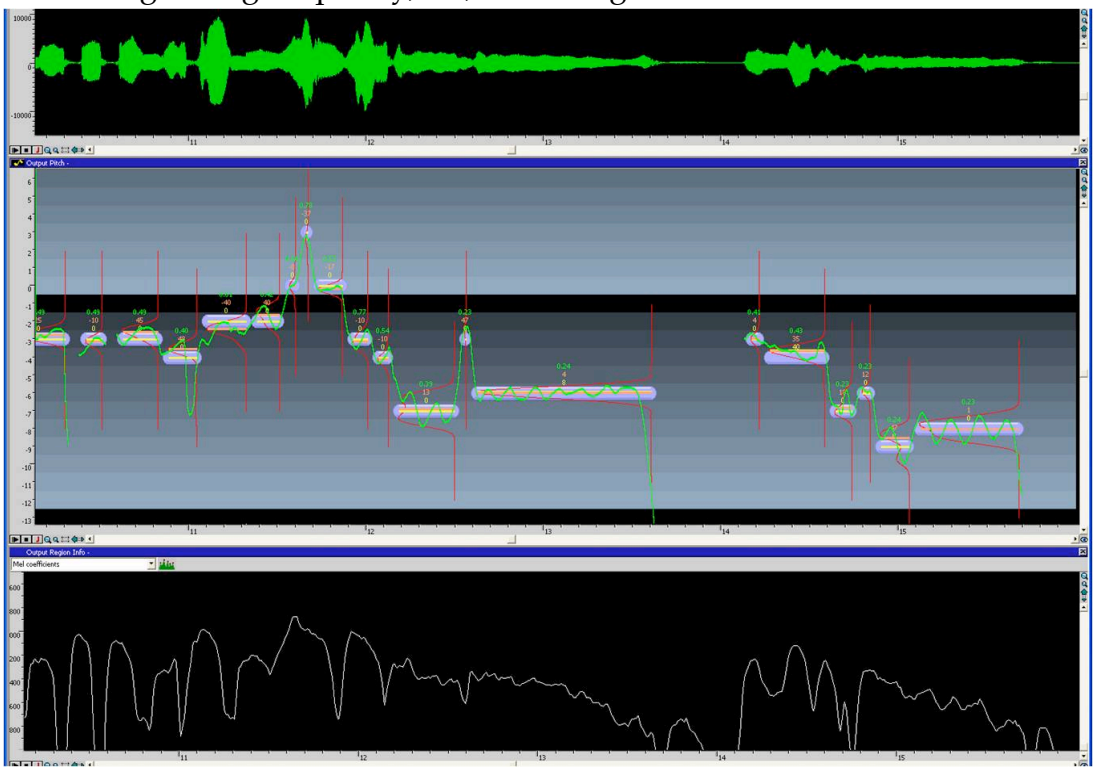

Figure 6: Example of the visualization tool for melodic transcription. Audio waveform (top), estimated f0 and pitch (middle) and energy (bottom).

Evaluation strategy

The comparison between automatic transcriptions and human annotations is very important for understanding the challenges of the task and the limitations of our methods. For evaluation, we gathered manual transcriptions from a representative music collection and considered standard evaluation metrics, as detailed below.

Music Collection

For this study, we have gathered a music collection of 72 sung excerpts representative of different a cappella singing styles (*Tonás*). This collection was built in the context of a study on similarity and style classification of flamenco a cappella singing styles. We refer to (Mora et al. 2010) for a comprehensive

description of the considered styles and their musical characteristics. All 72 excerpts are monophonic, their average duration is 30 seconds and there is enough variability for a proper evaluation of our methods, including a variety of singers, recording conditions, presence of percussion, clapping, background voices and noise. The files were manually transcribed to generate the ground truth. They contain a total of 211047 frames and 2803 notes, according to the ground truth.

In addition, we built a small control data set of pop/jazz a cappella singing, consisting of 5 musical phrases by different singers, recorded in good conditions. This control dataset will serve us to evaluate the difficulty of the flamenco material and test the algorithms in easier conditions to establish a performance ceiling.

Ground truth gathering (manual annotations)

As mentioned in Section 1, gathering manual transcriptions from flamenco scholars is very difficult, given the lack of standard methodology and the subjectivity of the task. This fact influences the evaluation of the computed transcriptions. We now explain the followed evaluation strategy.

Our goal was to assess our approach, which generates detailed transcriptions, i.e. including all the perceived notes, including ornamentations. Three subjects participated in the evaluation process: a musician with limited knowledge of flamenco music and two experts in flamenco music. The musician first carried out detailed manual annotations. As her knowledge in flamenco music was very limited, we expected her not to use implicit knowledge on the style. Annotations were then verified and corrected by a flamenco expert and occasionally discussed with another flamenco expert. In order to gather manual annotations, we provided the subjects with a user interface for visualizing the waveform and fundamental frequency (C*orrected-f0*) in cents (in a piano roll representation), as shown in Figure 6. As transcribing everything from scratch was very time

consuming, we also provided the subject with the output of a baseline transcription based on manually corrected fundamental frequency estimates (*Corrected-f0*). Please note that as the baseline transcription (*Corrected-f0)* is based on *sac* algorithm, this gives a slightly advantage to this method as compared to alternative ones. Subjects could listen to the waveform and the synthesized transcription (in MIDI-like format), while editing the melodic data (i.e. pitch, onset and offset of each note) until they were satisfied with the transcription. We observed that there was still a degree of subjectivity regarding the differentiation between ornaments and pitch glides, so subjects had to agree on and review their respective annotations.

Evaluation measures

We computed different evaluation measures as considered by (MIREX) in two related tasks: *Audio Melody Extraction*, consisting of the comparison of frame-based f0 values and pitch values to the ground truth; and *Multiple Fundamental Frequency Estimation & Tracking*, consisting of the comparison of note pitch and duration to the ground truth.

Regarding voicing, we consider two measures as stipulated by (MIREX): voicing recall, i.e. percentage of voiced frames according to the reference that are declared as voiced by the algorithm; and voicing false alarm, i.e. percentage of unvoiced frames according to the reference that are declared as voiced by the algorithm. Regarding pitch accuracy, we first considered raw pitch accuracy, i.e. percentage of voiced frames where the pitch estimation is correct, considering a certain tolerance or threshold in cents (*th*). This threshold is needed given the fact that frequency values are quantized to (equal-tempered) pitch values with respect to the estimated tuning frequency. This generates small mistunings of the estimated f0 envelopes. We have evaluated different threshold values.

We also consider raw chroma accuracy, defined as the percentage of voiced frames where the chroma estimation is correct, considering a certain tolerance or

threshold in cents (*th*). This measure allows for octave error in the estimation. Finally, overall accuracy represents the percentage of frames that have been correctly estimated in terms of pitch (for voiced frames) or correctly detected as unvoiced frames.

Regarding note accuracy, we compute the average note precision (ratio of correctly transcribed ground truth notes to the number of transcribed notes), recall (ratio of correctly transcribed ground truth notes to the number of ground truth notes) and f-measure. Here, a transcribed note is considered correct if its onset deviation is within 50 ms of a ground-truth note onset and its f0 is within +- 50 cents of the corresponding ground-truth note. In addition, a correctly estimated note is required to have a time offset deviation within 20% of the ground truth note duration, or within 50 ms whichever is larger. One ground truth note can only be associated with one transcribed note.

Finally, we have studied the influence in the evaluation results of two main steps of the algorithm. First, we analyzed the effect of the f0 estimation method by comparing several algorithms and a manually edited f0 envelopes, as described above. Second, we considered the influence of note segmentation by comparing our approach with an alternative one (*mami*) proposed by Mulder et al. (2003). Here, we didn't have access to the f0 envelope, but only to the final note-level melodic transcriptions.

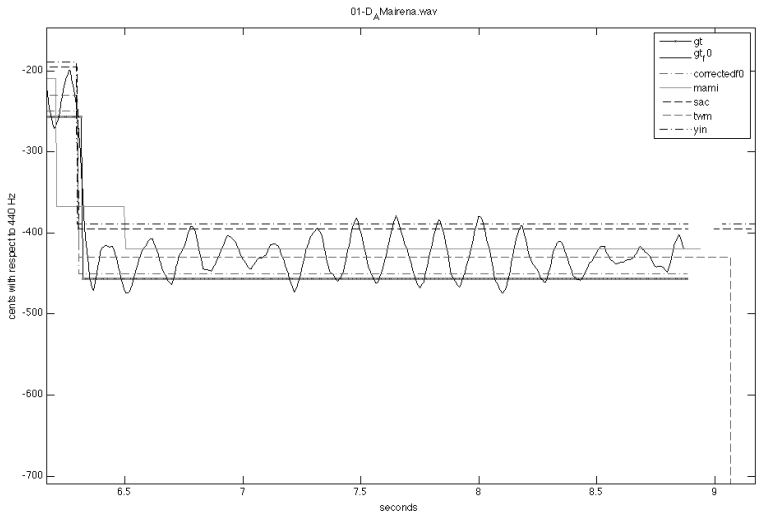

Figure 7: Examples of frame-based note estimation together with the fundamental frequency envelope (*Corrected-f0*) for a single note in an excerpt from a Debla style (by singer Antonio Mairena).

Results

This section summarizes the qualitative and quantitative results of the different approaches, including an error analysis and a brief summary on how the automatic transcription system has been used in different application contexts.

Examples and observations

We analyze here the f0 and note transcription outputs for the different methods. We first observe that there is a small detuning among the different outputs, as can be seen in Figure 7, because of the following reason. As mentioned above, the algorithm for note segmentation estimates the tuning frequency and then quantizes the pitch values according to an equal-tempered scale using the obtained tuning frequency. The fact that the tuning frequency is estimated according to the note values means that different approaches yield different values for the tuning frequency. This then generates small transpositions (usually less than a semitone) between the estimated transcriptions. In addition, the algorithm assumes a constant tuning with respect

to 440 Hz. We observe that, for some excerpts, the singer changes the tuning along time, meaning our assumption does not hold. This also creates detuning in the transcription. For that reason, it is important to compare the obtained representations considering this small detuning (by means of the tolerance parameter *th*).

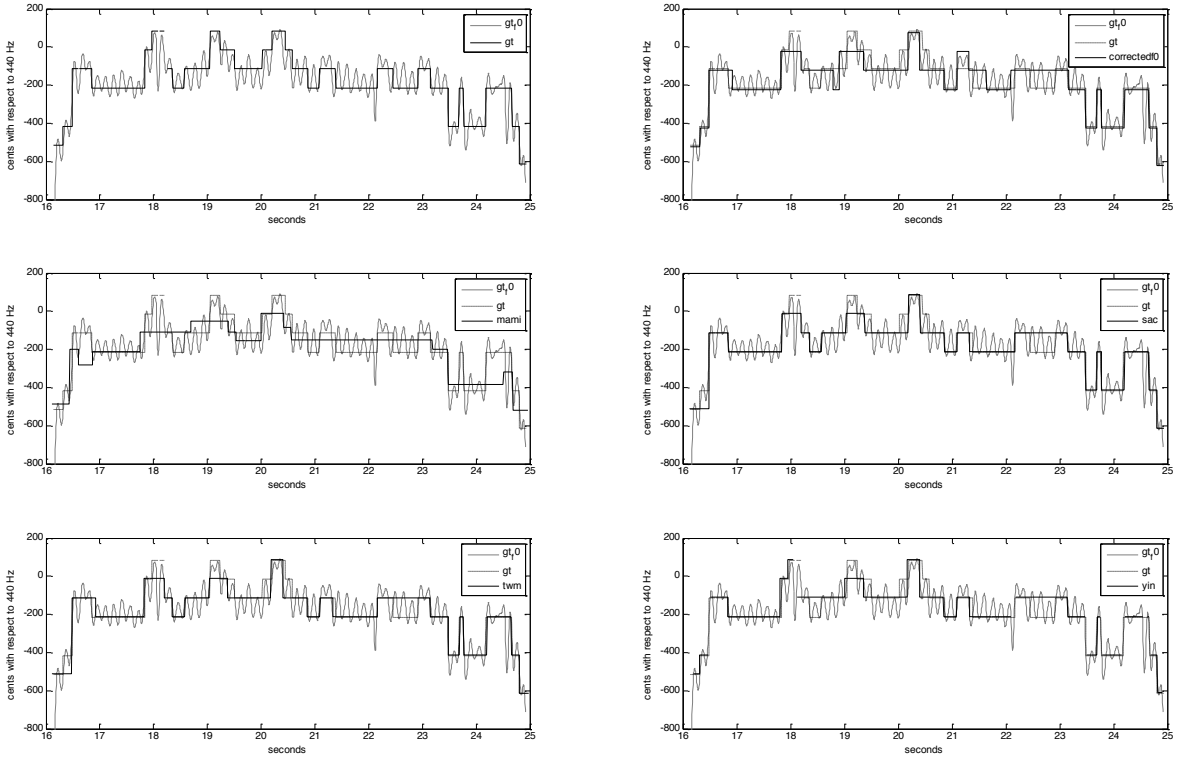

Figure 8: Examples of frame-based note transcription together with the fundamental frequency envelope (*Corrected-f0*) in an excerpt in the Debla style by singer Naranjito.

We also see that different algorithms produce different segmentation results in short notes, as illustrated in Figure 8. This is due to the fact that the note consolidation procedure highly depends on the input f0 envelope, which varies between algorithms, as each does its own fundamental frequency detection.

Frame-based accuracy

The evaluation results for different tolerance intervals are provided in Figure 9. When considering *th*=100 cents (1 semitone), the segmentation algorithm proposed in this study yields the best overall accuracy when using the corrected

f0 envelope (90.43%), followed by the proposed note segmentation approach and *sac* (81.68%) and *twm* (79.14%) f0 estimation algorithms. Results for the evaluated state-of-the-art approach (*mami)* are also very close to our system (79.1%).

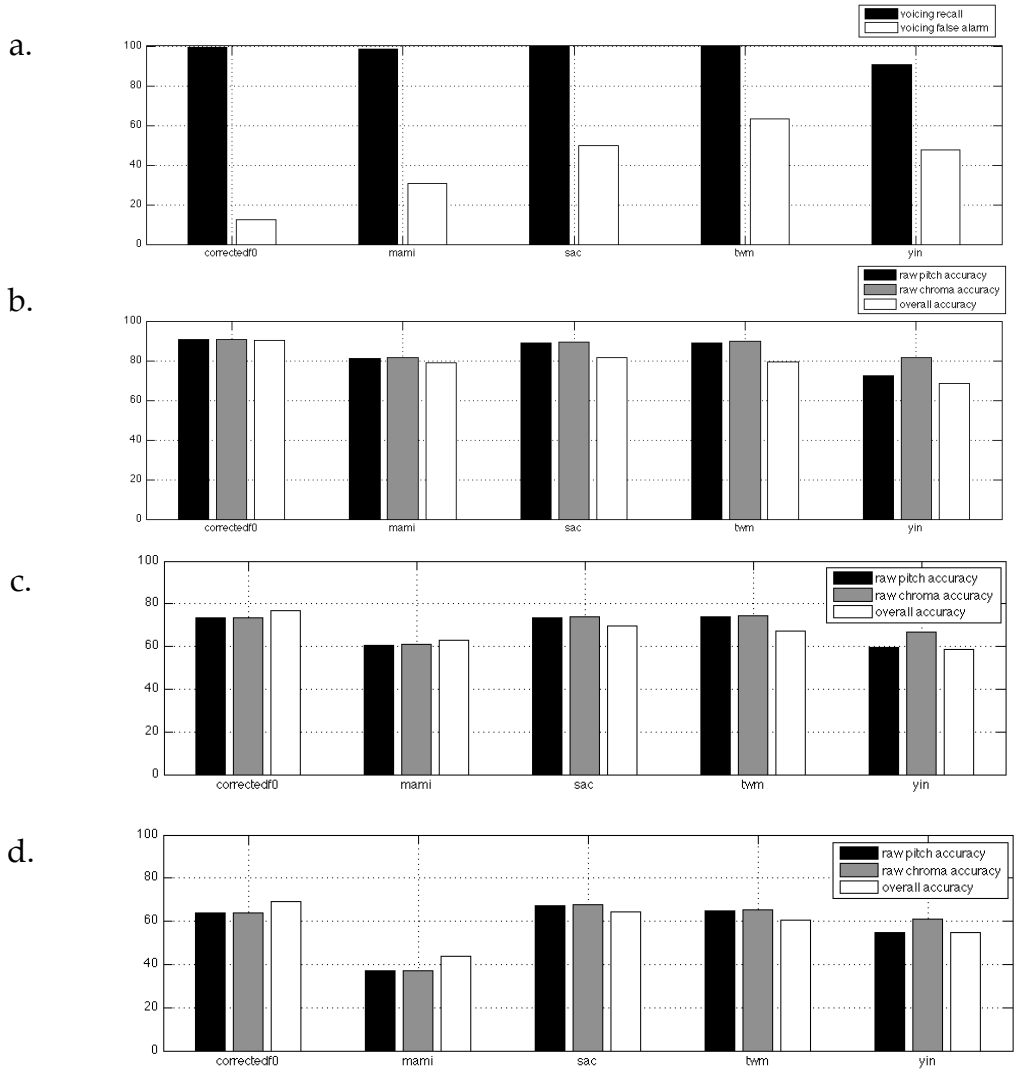

Figure 9: Frame-based accuracy measures: voicing recall and voicing false alarm (a), where the voicing recall rate should be maximized whilst the voicing false alarm rate should be minimized; accuracy measures with *th*=100 cents (b), *th*=50 cents (c) and *th*=25 cents (d).

The worst results are yielded by the *yin* algorithm (68.56%). In terms of pitch accuracy, the proposed approach (using *sac* or *twm* f0 estimation algorithms) outperforms *mami* and *yin*. We believe this is probably due to the fact that both *sac* and *twm* algorithms have been especially designed for singing voice and that

manual transcriptions might provide a slightly advantage to *sac* as compared to alternative methods, as we already discussed.

We also observe that *mami* has better (i.e. lower) voicing false alarm rate (*vx_false_alm_av*) than our approach. This fact together with the high difference in voicing false alarm between *Corrected-f0* and *sac* or *twm* indicates that the system would benefit from an improved voicing detection procedure after f0 estimation.

If we decrease the tolerance to half a semitone (*th*=50 cents), the overall accuracy decreases for all the considered approaches (e.g. 76.81% for *Corrected-f0*, 69.56% for *sac*). The ranking of the algorithms is also similar to the 100 cents tolerance, although the accuracy of the *mami* approach is closer to *yin*. Finally, for 25 cents tolerance, the ranking of methods is almost the same but with the difference that *mami* gets the lowest *overall accuracy*. This might be due to the fact that *mami* does not quantize note pitches to an equal-tempered scale as it is done in the ground truth. The overall accuracy with decreasing tolerance is shown in Figure 10. As expected, the overall accuracy increases with the tolerance for all the considered approaches.

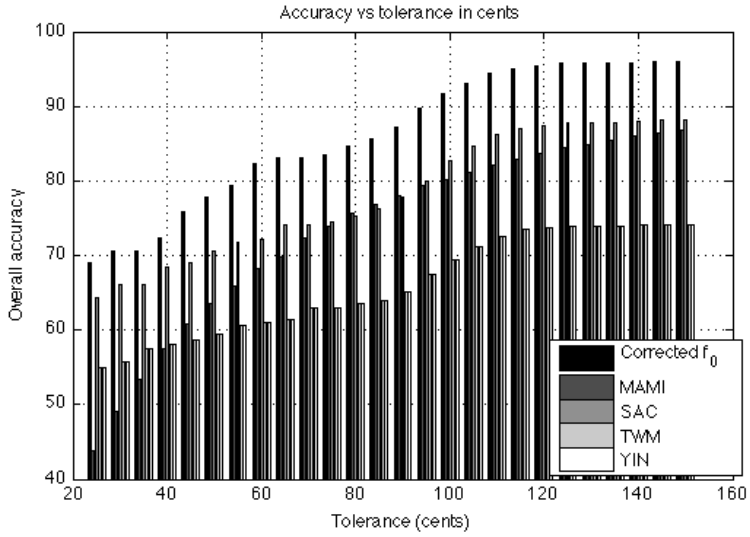

Figure 10: Overall accuracy with respect to tolerance (*th* ∈ [25, 150] cents).

We also provide in Figure 11 a comparison of the accuracy for flamenco vs rock/jazz singing. We observe that the highest overall accuracy obtained for

pop/jazz singing is equal to 87%, and, as expected, it supports our hypothesis that all the methods work better for these singing styles. As a general conclusion, we observe that having good f0 estimations is crucial for melodic transcription.

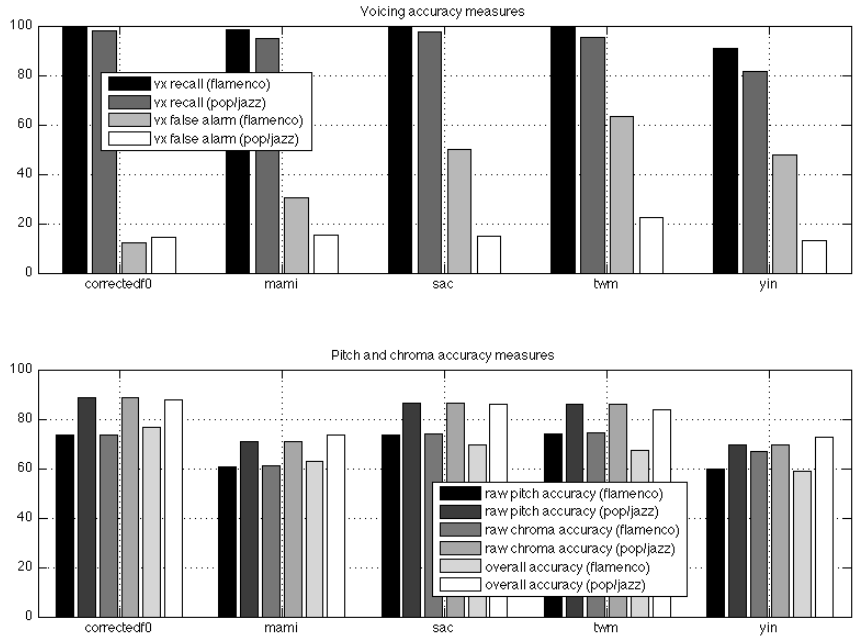

Figure 11: Comparison of frame-based accuracy measures (50 cents tolerance) for flamenco against a control dataset from pop/jazz. The voicing recall rate should be maximized whilst the voicing false alarm rate should be minimized

Note Accuracy

Figure 12(a) shows the note accuracy measures, precision, recall and f-measure, for all the considered approaches. Black bars correspond to the results obtained with the whole transcription system, including the tuning frequency iterative refinement. We first observe that the highest f-measure is very low, 32%, and this is due to a very strict design of the evaluation measure. In fact, it is the same order of magnitude than current MIREX results (the best f-measure for the Multiple Note Tracking was 34.9% in 2011) (MIREX). Although MIREX deals with very simple polyphonies, we have previously stated the challenges of our material: noise, ornamentation, detuning, continuous variation of energy and f0

and difficulty of agreement among experts. In this measure, for instance, if one ornament is not correctly detected (e.g. three notes in the ground truth consolidated in the transcription), none of the notes from the ground truth or the transcription are added up in precision and recall, as their onset and duration do not fulfil the requirements. In addition, notes are not weighted by their duration. Moreover, we observe than our approach outperforms *mami*, which yields an f-measure equal to 15%. This result is quite different from the results reported by Mulder et al. (2003) against alternative systems for transcribing sung queries: singing with syllables (414 notes), with words (657 notes) and whistling (283 notes), obtaining error rates between 10% and 20%. There are indeed few details on the way this error is computed, although we observe that the evaluation measure is much more tolerant (they indicate that '*total error is obtained by adding the percent of times a MIDI- rounded note difference of 2 or more semitones is observed'*). In addition, we observe similar algorithm ranking as for frame-based measures, i.e. *sac* and our note segmentation method outperforms alternative f0 estimation algorithms and *mami*.

In the same figure, white bars represent the results obtained when we disable the tuning frequency refinement procedure in the note transcription process. The performance decays as expected, showing the benefits of tuning frequency refinement, through the differences are not considerable: f-measure decays 6% for *sac*, 3% for *twm*, and 4% for *yin*.

If we increase the tolerance of our evaluation measures (150 ms for onsets, 30% for durations, and 75 cents for pitch), the f-measure increases to almost 40% as shown in Figure 12(b). In any case, the results are much higher for the control pop/jazz dataset for all the considered approaches, with a maximum f-measure of 62%. This again reveals the complexity of dealing with our particular material.

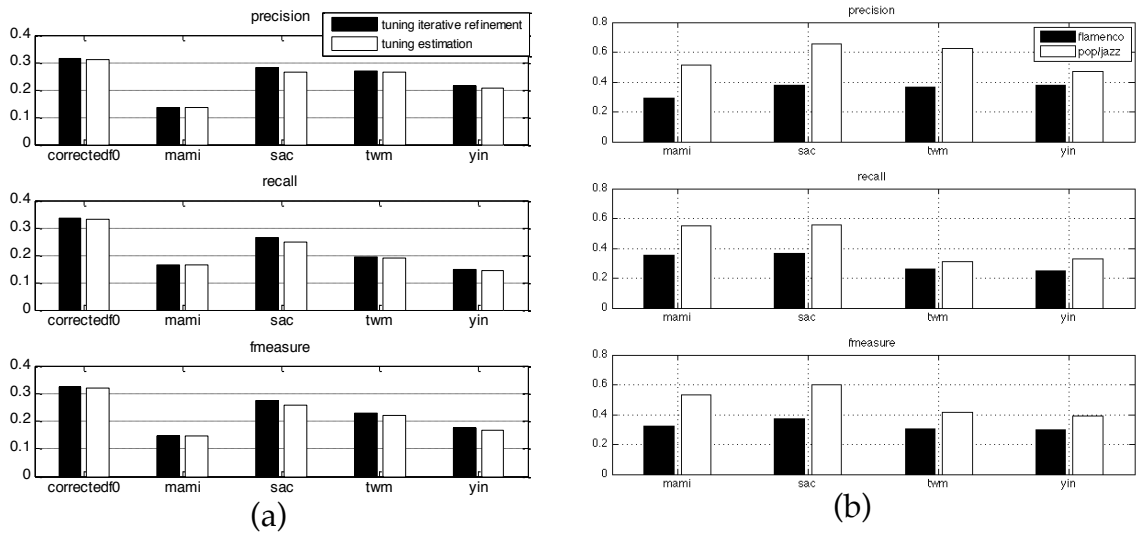

Figure 12: Note accuracy measures for the considered approaches. (a) Influence of the tuning frequency refinement procedure. (b) Flamenco vs pop/jazz testset.

Error Analysis

We observe that for some of the excerpts most algorithms fail, due to three main reasons. The first reason, mentioned before, is that tuning might be variable over time, which contradicts the assumption of constant tuning. Second, the different algorithms do not often correctly segment short notes; either they are consolidated while the annotation consists of several close notes, or vice versa. This especially happens in highly ornamented passages with unstable f0 contour. Third, the f0 estimation fails on noisy passages or with a high degree of reverberation (specially at the end of phrases). Finally ,we found that the *yin* algorithm had some octave errors in several of the analyzed excerpts, which generated errors in the note segmentation procedure.

Validation in Different Application Contexts

As an additional validation phase, the extracted transcriptions have been used in two different application contexts. The first one analyzes the overall melodic contour, which is used to measure melodic similarity as a way to characterize and classify a capella singing styles and variants. In this context, we

performed some post-processing of the melodic transcription, including the conversion from pitch to interval representation and the simplification of the transcriptions by consolidating short notes. The reader is referred to (Cabrera et al. 2008; Mora et al. 2010) for more details on this application.

The second application context analyzes the detailed annotations to detect frequent and representative ornamentation (*melisma*) in flamenco singing. It is based on strategies for pattern detection which consider both instantanous f0 values and note pitch and duration information (Gomez et al. 2011). Here, the computed transcriptions were only post-processed in order to convert pitch to interval values.

Conclusions and Future Perspectives

This paper proposes an approach for automatic transcription of flamenco a cappella singing. We have analyzed the main technological challenges, and proposed an approach based on and iterative note segmentation and labelling technique from f0, energy and timbre. The approach has been evaluated for a collection of annotated performances, obtaining satisfactory results for different f0 estimation algorithms (*twm* and *sac*), comparable to a state-of-the-art approach (*mami*). We also observed that there are still some limitations on noisy passages or with a high degree of reverberation (errors in f0 estimation), singers with bad tuning (wrong note pitches) and in highly ornamented passages with unstable f0 contour (wrong note segmentations).

Our approach has been successfully validated for comparing performances, styles and variants by means of melodic similarity and for locating frequent ornamentation. This validation has been carried out in the context of a research project in collaboration with flamenco experts (COFLA: Computational Analysis of Flamenco Music http://mtg.upf.edu/research/projects/cofla). Moreover, we gathered some positive feedback from flamenco experts outside of our group during the IIIrd Interdisciplinary Congress on Flamenco Research (INFLA 2012).

In the future, we plan to carry out a proper user study to investigate how flamenco experts can benefit from the system.

We have seen that there is still much room for improvement, given the difficulty of the task. As future work we intend to address the mentioned system limitations, extend the amount of manual annotations, compare annotations of independent experts as a way to quantify the incertainty of the ground truth information, and implement an adaptive strategy to make the algorithm learn from user annotations.

Finally, the availability of predominant f0 estimation algorithms has allowed us to extend this approach to polyphonic music signals, mainly solo voice with guitar accompaniment, by integrating state-of-the-art predominant f0 estimation techniques (Gomez et al. 2012).

Acknowledgements

The authors would like to thank the COFLA team[i] for providing the data set and expert knowledge in flamenco music. We also thank Micheline Lesaffre and authors of (Mulder et al. 2003) for granting access to the mami executable. This work has been partially funded by AGAUR BE-DGR 2009 B (post-doctoral mobility grant) and the COFLA project (P09-TIC-4840 Proyecto de Excelencia, Junta de Andalucía).

References

Blas Vega, J., & Ríos Ruiz, M. (1988) Diccionario enciclopédico ilustrado del flamenco. Cinterco. Madrid.

Cabrera, J.J., Díaz-Bañez, J.M., Escobar-Borrego, F.J., Gómez, E., & Mora, J. (2008). *Comparative Melodic Analysis of A Cappella Flamenco Cantes*. Proceedings of Conference on Interdisciplinary Musicology, pp. 38-39.

Cano, P. (1998). Fundamental Frequency Estimation in the SMS analysis. Proceedings of International Conference on Digital Audio Effects (DAFX), pp. 99-102.

Cardine, E. (1970). Sémiologie grégorienne. Etudes Grégoriennes. Abbaye Saint Pierre de Solesmes, Sablé sur Sarthe, Tome XI.

De Cheveigné, A., & Kawahara, H. (2002). *YIN, a fundamental frequency estimator for speech and music*. Journal of the Acoustical Society of America, 111, pp. 1917-1930.

Donnier, P. (1996*). La musique flamenca: composition et structure.* PhD thesis, Université Paris X-Nanterre *(Département d'ethnologie et de sociologie comparative).*

Donnier, P. (2011). *Flamenco: elementos para la transcripción del cante y la guitarra*. Gómez Martín, F. (ed) Sección Música y Matemáticas, Revista digital Divulgamat, Real Sociedad Matemática Española, April 2011. http://divulgamat2.ehu.es/divulgamat15/index.php?option=com_content&view=article&id=12354&directory=67.

Fernández, L. (2004). *Teoría musical del flamenco*. Acordes concert. Madrid.

Gamboa, J. M. (2005) *Una historia del flamenco*. Espasa-Calpe. Madrid.

Gedik, A. C., & Bozkurt, B. (2010) *Pitch Frequency Histogram Based Music Information Retrieval for Turkish Music*, Signal Processing, vol.10, pp.1049-1063. (doi:10.106/j.sigpro.2009.06.017)

Gómez, E., Klapuri, A., & Meudic, B. (2003). *Melody Description and Extraction in the Context of Music Content Processing*, Journal of New Music Research, vol.32 (1), pp. 23-40.

Gómez, F., Pikrakis, A., Mora, J., Díaz-Báñez, J. M., Gómez, E., & Escobar, F. (2011) *Automatic detection of ornamentation in flamenco music*, Proceedings of International Workshop on Music and Machine Learning: Learning from Musical Structure, Neural Information Processing Systems Foundation, Granada, December 2011.

Gómez, E., Cañadas F., Salamon J., Bonada J., Vera P., & Cabañas P. (2012). *Predominant Fundamental Frequency Estimation vs Singing Voice Separation for*

the Automatic Transcription of Accompanied Flamenco Singing. Proceedings of International Society for Music Information Retrieval Conference, pp. 601-606.

Hoces, R. (2011) *La transcripción musical para guitarra flamenca: análisis e implementación metodológica*. PhD Thesis, Universidad de Sevilla.

Hurtado Torres, D., & Hurtado Torres A. (1998) *El arte de la escritura musical flamenca*. Bienal de Arte Flamenco. Sevilla.

Hurtado Torres, A., & Hurtado Torres, D. (2002). *La voz de la tierra, estudio y transcripción de los cantes campesinos en las provincias de Jaén y Córdoba*. Centro Andaluz de Flamenco. Jerez.

Janer, J., Bonada, J., de Boer, M., & Loscos, A. (2008). *Audio Recording Analysis and Rating*, Patent pending US20080026977, Universitat Pompeu Fabra, 06/02/2008.

Katz, Israel J. *Flamenco*. Grove Music Online ed. L. Macy (Accessed 16 May, 2006), www.grovemusic.com

Klapuri, A., & Davy, M. (Editors) (2006). *Signal Processing Methods for Music Transcription*. Springer-Verlag, New York.

Lesaffre, M., Leman, M., De Baets, B., & Martens, J.-P. (2004). *Methodological considerations concerning manual annotation of musical audio in function of algorithm development*, Proceedings of the International Conference on Music Information Retrieval, pp. 64-71.

Maher, R. C., & Beauchamp, J. W. (1994). *Fundamental frequency estimation of musical signals using a two-way mismatch procedure*, Journal of the Acoustical Society of America, Vol. 95(4), pp. 2254-2263.

MIREX wiki, http://www.music-ir.org/mirex/wiki/, accessed November, 14, 2011.

Mora, J., Gomez, F., Gomez, E., Escobar-Borrego, F.J., & Diaz-Bañez, J.M. (2010). *Melodic Characterization and Similarity in A Cappella Flamenco Cantes*.

Proceedings of the International Society for Music Information Retrieval Conference, pp. 351-356.

Mulder, T., Martens, J. P. Lesaffre, M., Leman, M., De Baets, B., & De Meyer, H. (2003), *An Auditory Model Based Transcriber of Vocal Queries*, Proceedings of the International Conference on Music Information Retrieval, pp. 26-30.

Navarro, J.L., & Ropero, M. (editors) (1995). *Historia del flamenco*. Ed. Tartessos, Sevilla.

Ness, S. R., Biró, D. P., & Tzanetakis, G. (2010). *Computer-assisted cantillation and chant research using content-aware web visualization tools*. Multimedia Tools Appl. 48(1), pp. 207-224.

Ryynänen, M. P. (2006). *Singing transcription*, in *Signal processing methods for music transcription* (A. Klapuri and M. Davy, eds.), Springer, New York, pp. 361-390.

Six, J., & Cornelis, O. (2011) *Tarsos - a Platform to Explore Pitch Scales in Non-Western and Western Music*. Proceedings of the International Conference on Music Information Retrieval, pp. 169-174.

Sundberg, J. (1987). *The Science of the Singing Voice*. DeKalb, IL: Northern Illinois Univ. Press.

Toiviainen, P. & Eerola, T. (2006). *Visualization in comparative music research*. In A. Rizzi & M Vichi (Eds.), COMPSTAT 2006 - Computational Statistics. Heidelberg: Physica-Verlag, pp. 209-221.

Vidwans, A. & Rao, P. *Identifying Indian Classical Music Styles using Melodic Contours*, Proceedings of Frontiers of Research on Speech and Music (FRSM), January 2012, Gurgaon, India.