EVALUATION FRAMEWORK FOR AUTOMATIC SINGING TRANSCRIPTION

Emilio Molina, Ana M. Barbancho, Lorenzo J. Tardón, Isabel Barbancho
Universidad de Málaga, ATIC Research Group, Andalucía Tech,
ETSI Telecomunicación, Campus de Teatinos s/n, 29071 Málaga, SPAIN
emm@ic.uma.es, abp@ic.uma.es, lorenzo@ic.uma.es, ibp@ic.uma.es

ABSTRACT

In this paper, we analyse the evaluation strategies used in previous works on automatic singing transcription, and we present a novel, comprehensive and freely available evaluation framework for automatic singing transcription. This framework consists of a cross-annotated dataset and a set of extended evaluation measures, which are integrated in a Matlab toolbox. The presented evaluation measures are based on standard MIREX note-tracking measures, but they provide extra information about the type of errors made by the singing transcriber. Finally, a practical case of use is presented, in which the evaluation framework has been used to perform a comparison in detail of several state-of-the-art singing transcribers.

1. INTRODUCTION

Singing transcription refers to the automatic conversion of a recorded singing signal into a symbolic representation (e.g. a MIDI file) by applying signal-processing methods [1]. One of its renowned applications is query-by-humming [5], but other types of applications also are related to this task, like singing tutors [2], computer games (e.g. Singstar [1]), etc. In general, singing transcription is considered a specific case of melody transcription (also called note tracking), which is more general problem. However, singing transcription not only relates to melody transcription but also to speech recognition, and still nowadays it is a challenging problem even in the case of monophonic signals without accompaniment [3].

In the literature, various approaches for singing transcription can be found. A simple but commonly referenced approach was proposed by McNab in 1996 [4], and it relied on several handcrafted pitch-based and energy-based segmentation methods. Later, in 2001 Haus et al. used a similar approach with some rules to deal with intonation issues [5], and in 2002, Clarisse et al. [6] contributed with an auditory model, leading to later improved systems such as [7] (later included in MAMI project [2] and today in SampleSumo products [3]). Additionally, other more recent approaches use hidden Markov models (HMM) to detect note-events in singing voice [8, 9, 11]. One of the most representative HMM-based singing transcribers was published by Ryynänen in 2004 [9]. More recently, in 2013, another probabilistic approach for singing transcription has been proposed in [3], also leading to relevant results. Regarding the evaluation methodologies used in these works (see Sections 2.1 and 3.1 for a review), there is not a standard methodology.

In this paper, we present a comprehensive evaluation framework for singing transcription. This framework consists of a cross-annotated dataset (Section 2) and a novel, compact set of evaluation measures (Section 3), which report information about the type of errors made by the singing transcriber. These measures have been integrated in a freely available Matlab toolbox (see Section 3.3). Then, we present a practical case in which the evaluation framework has been used to perform a comparison in detail of several state-of-the-art singing transcribers (Section 4). Finally, some relevant conclusions are presented in Section 5

2. DATASETS

In this section, we review the evaluation datasets used in prior works on singing transcription , and we describe the proposed evaluation dataset and our strategy for ground-truth annotation.

2.1 Datasets used in prior works

In Table 1, we present the datasets used in some relevant works on singing transcription. Note that none of the datasets fully represents the possible contexts in which singing transcription might be applied, since they are either too small (e.g. [5,6]), either very specific in style (e.g. [11] for opera and [3] for flamenco), or either they use an annotation strategy that may be subjective (e.g. [5, 6]), or only valid for very good performances in rhythm and intonation (e.g. [8, 9]). In addition, only the flamenco dataset used in [3] is freely available.

2.2 Proposed dataset

In this section we describe the music collection, as well as the annotation strategy used to build the ground-truth.

[1] http://www.singstar.com

[2] http://www.ipem.ugent.be/MAMI

[3] http://www.samplesumo.com

Author	Year	Dataset size	Audio quality	Music style	Singing style	Ground-truth (GT) annotation strategy	Tunning devs. annotated in GT	Freely available
McNab [4]	1996	NONE						
Haus & Pollastri [5]	2001	20 short melodies	Low & moderated noise	Popular and scales	Syllables: 'na-na'...	Annotated by one musician	No	No
Clarisse et al. [6]	2002	22 short melodies	Low & moderated noise	Popular	Singing with & without lyrics	Annotation by one musician	No	No
Viitaniemi et al. [8]	2003	66 melodies (120 minutes)	High quality (studio conditions)	Folk songs & scales	Singing, humming & whistling	Original score used as ground-truth	No	No
Ryynänen et al. [9]	2004							
Mulder et al. [7]	2004	52 melo. (1354 notes)	Good & moderated noise	Popular songs	Syllables, singing & whistling	Team of musicologists	No	No
Kumar et al. [10]	2007	47 songs (2513 notes)	Good	Indian music	Syllables: /la/ /da/ /na/	Manual annot. of vowel onsets [REf]	No	No
Krige et al. [11]	2008	13842 notes	High quality but strong reverberation	Opera lessons & scales	Syllables	Time alignment using Viterbi	No	No
Gómez & Bonada [3]	2013	72 excerpts (2803 notes)	Good & slightly noisy	Flamenco songs	Lyrics & ornaments	Musicians team (cross-annotation)	Yes	Yes

Table 1. Review of the evaluation datasets used in prior works on singing transcription. Some details about the dataset are not provided in some cases, so certain fields can not be expressed in the same units (e.g. dataset size).

2.2.1 Music collection

The proposed dataset consists of 38 melodies sung by adult and child untrained singers, recorded in mono with a sample rate of 44100Hz and a resolution of 16 bits. Generally, the recordings are not clean and some background noise is present. The duration of the excerpts ranges from 15 to 86 seconds and the total duration of the whole dataset is 1154 seconds. This music collection can be broken down into three categories, according to the type of singer:

- Children (our own recordings [4]): 14 melodies of traditional children songs (557 seconds) sung by 8 different children (5-11 years old).
- Adult male: 13 pop melodies (315 seconds) sung by 8 different adult male untrained singers. These recordings were randomly chosen from the public dataset MTG-QBH [5] [12].
- Adult female: 11 pop melodies (281 seconds) sung by 5 different adult female untrained singers, also taken from MTG-QBH dataset.

Note that in this collection the pitch and the loudness can be unstable, and well performed vibratos are not frequent.

2.2.2 Ground-truth: annotation strategy

The described music collection has been manually annotated to build the ground truth [4] . First, we have transcribed the audio recordings with a baseline algorithm (Section 4.2), and then all the transcription errors have been corrected by an expert musician with more than 10 years of music training. Then, a second expert musician (with 7 years of music training) checked all the annotations until both musicians agreed in their correctness. The transcription errors were corrected by listening, at the same time, to the synthesized transcription and the original audio. The musicians were given a set of instructions about the specific criteria to annotate the singing melody:

- Ornaments such as pitch bending at the beginning of the notes or vibratos are not considered independent notes. This criterion is based on Vocaloid's [6] approach, where ornaments are not modelled with extra notes.
- Portamento between two notes does not produce an extra third note (again, this is the criteria used in Vocaloid).
- The onsets are placed at the beginning of voiced segments and in each clear change of pitch or phoneme. In the case of 'l', 'm', 'n' voiced consonants + vowel (e.g. 'la'), the onset is not placed at the beginning of the consonant but at the beginning of the vowel.
- The pitch of each note is annotated with cents resolution as perceived by the team of experts. Note that we annotate the tuning deviation for each independent note.

3. EVALUATION MEASURES

In this section, we describe the evaluation measures used in prior works on automatic singing transcription, and we present the proposed ones.

3.1 Evaluation measures used in prior works

In Table 2, we review the evaluation measures used in some relevant works on singing transcription. In some cases, only the note and/or frame error is provided as a compact, representative measure [5, 9], whereas other approaches provide extra information about the type of errors made by the system using dynamic time warping (DTW) [6] or Viterbi-based alignment [11]. In our case, we have taken the most relevant aspects of these approaches and we added some novel ideas in order to define a novel, compact and comprehensive set of evaluations.

[4] Available at http://www.atic.uma.es/ismir2014singing

[5] http://mtg.upf.edu/download/datasets/mtg-qbh

[6] http://www.vocaloid.com

Author	Year	Evaluation measures
McNab	1996	NONE
Haus & Pollastri [5]	2001	Rate of note pitch errors (segmentation errors are not considered)
Clarisse et al. [6]	2002	DTW-based measurement of various note errors, e.g. insertions deletions and substitutions.
Viitaniemi et al. [8]	2003	Frame-based errors. Do not report information about type of errors made.
Ryynänen et al. [9]	2004	Note-based and frame-based errors. Do not report information about type of errors made.
Mulder et al. [7]	2004	DTW-based measurement of various note errors, e.g. insertions deletions and substitutions.
Kumar et al. [10]	2007	Onset detection errors (pitch and durations are ignored).
Krige et al. [11]	2008	Viterbi-based measurement of deletions, insertions and substitutions (typical evaluation in speech recognition).
Gómez & Bonada [3]	2013	MIREX measures for audio melody extraction and note-tracking. Do not report information about type of errors made.

Table 2. Evaluation measures used in prior works on singing transcription.

3.2 Proposed measures

In this section, we firstly present the notation and some needed definitions that are used in the rest of sections, and then we describe the evaluation measures used to quantify the proportion of correctly transcribed notes. Finally, we present a set of novel evaluation measures that independently report the importance of each type of error. In Figure 1 we show an example of the types of errors considered.

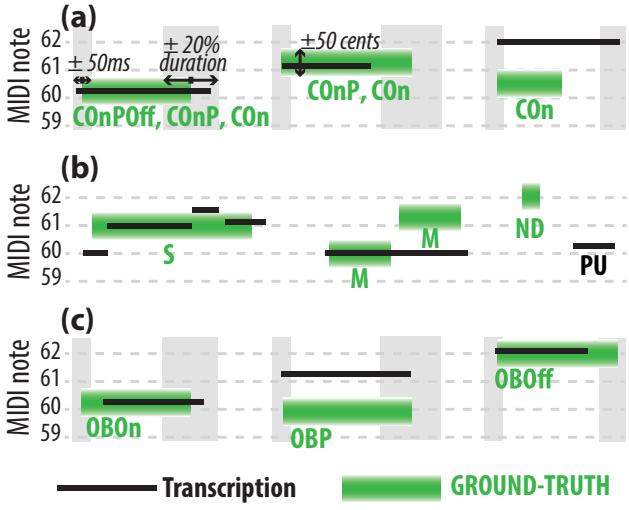

Figure 1. Examples of the different proposed measures.

3.2.1 Notation

The i:th note of the ground-truth is noted as n_i^{GT}, and the j:th note of the transcription is noted as n_j^{TR}. The total number of notes in the ground-truth and the transcription are N^{GT} and N^{TR}, respectively. Regarding the expressions used in the for correct notes, we have used Precision, Recall and F-measure, which are defined as follow:

$$\mathbf{CX}_{\text{Precision}} = \frac{N_{\mathbf{CX}}^{\text{GT}}}{N^{\text{GT}}} \quad (1)$$

$$\mathbf{CX}_{\text{Recall}} = \frac{N_{\mathbf{CX}}^{\text{TR}}}{N^{\text{TR}}} \quad (2)$$

$$\mathbf{CX}_{\text{F-measure}} = 2 \cdot \frac{\mathbf{CX}_{\text{Precision}} \cdot \mathbf{CX}_{\text{Recall}}}{\mathbf{CX}_{\text{Precision}} + \mathbf{CX}_{\text{Recall}}} \quad (3)$$

where $C\mathbf{X}$ makes reference to the specific category of correct note: Correct Onset & Pitch & Offset ($\mathbf{X}$ = COnPOff), Correct Onset & Pitch ($\mathbf{X}$ = COnP) or Correct Onset ($\mathbf{X}$ = COn). Finally, $N_{\mathbf{CX}}^{\text{GT}}$ and $N_{\mathbf{CX}}^{\text{TR}}$ are the total number of matching $C\mathbf{X}$ conditions in the ground-truth and the transcription, respectively.

Regarding the measures used for errors, we have computed the Error Rate with respect to N^{GT}, or with respect to N^{TR}, as follow:

$$\mathbf{X}_{\text{RateGT}} = \frac{N_{\mathbf{X}}^{\text{GT}}}{N^{\text{GT}}} \quad (4)$$

$$\mathbf{X}_{\text{RateTR}} = \frac{N_{\mathbf{X}}^{\text{TR}}}{N^{\text{TR}}} \quad (5)$$

Finally, in the case of segmentation errors (Section 3.2.5), we also compute the mean number of notes tagged as $\mathbf{X}$ in the transcription for each note tagged as $\mathbf{X}$ in the ground-truth. This magnitude has been expressed as a ratio:

$$\mathbf{X}_{\text{Ratio}} = \frac{N_{\mathbf{X}}^{\text{TR}}}{N_{\mathbf{X}}^{\text{GT}}} \quad (6)$$

3.2.2 Definition of correct onset/pitch/offset

The definitions of correctly transcribed notes (given in Section 3.2.3) consists of combinations of three independent conditions: correct onset, correct pitch and correct offset. We have defined these conditions according to MIREX (*Multiple F0 estimation and tracking* and *Audio Onset Detection* tasks), and so they are defined as follow:

• Correct Onset: If the note's onset of a transcribed note n_j^{TR} is within a $\pm$50ms range of the onset of a ground-truth note n_i^{GT}, i.e.:

$$\text{onset}(n_j^{TR}) \in [\text{onset}(n_i^{GT}) - 50\text{ms}, \text{onset}(n_i^{GT}) + 50\text{ms}] \quad (7)$$

then we consider that n_i^{GT} has a correct onset with respect to n_j^{TR}.

• Correct Pitch: If the note's pitch of a transcribed note n_j^{TR} is within a ± 0.5 semitones range of the pitch of a ground-truth note n_i^{GT}, i.e.:

$$\text{pitch}(n_j^{TR}) \in [\text{pitch}(n_i^{GT}) - 0.5\text{ st}, \text{pitch}(n_i^{GT}) + 0.5\text{ st}] \quad (8)$$

then we consider that n_i^{GT} has a correct pitch with respect to n_j^{TR}.

• Correct Offset: If the offsets of the ground-truth note n_i^{GT} and the transcribed note n_j^{TR} are within a range of $\pm 20\%$ of the duration of n_i^{GT} or ± 50 ms, whichever is larger, i.e.:

$$\text{offset}(n_j^{TR}) \in [\text{offset}(n_i^{GT}) - \text{OffRan}, \text{offset}(n_i^{GT}) + \text{OffRan}] \quad (9)$$

where $\text{OffRan} = max(50\text{ms}, \text{duration}(n_i^{GT}))$, then we consider that n_i^{GT} has a correct offset with respect to n_j^{TR}.

3.2.3 Correctly transcribed notes

The definition of "correct note" should be useful to measure the suitability of a given singing transcriber for a specific application. However, different applications may require a different definition of correct note. Therefore, we have chosen three different definitions of correct note as defined in MIREX:

• Correct onset, pitch and offset (COnPOff): This is a standard correctness criteria, since it is used in MIREX (*Multiple F0 estimation and tracking* task), and it is the most restrictive one. The note n_i^{GT} is assumed to be correctly transcribed into the note n_j^{TR} if it has correct onset, correct pitch and correct offset (as defined in Section 3.2.2). In addition, one ground truth note n_i^{GT} can only be associated with one transcribed note n_j^{TR}. In our evaluation framework, we report Precision, Recall and F-measure as defined in Section 3.2.1:

$$\text{COnPOff}_{\text{Precision}}, \text{COnPOff}_{\text{Recall}} \text{ and } \text{COnPOff}_{\text{F-measure}}.$$

• Correct Onset, Pitch (COnP): This criteria is also used in MIREX, but it is less restrictive since it just considers onset and pitch, and ignores the offset value. Therefore, in COnP criteria, a note n_i^{GT} is assumed to be correctly transcribed into the note n_j^{TR} if it has correct onset and correct pitch. In addition, one ground truth note n_i^{GT} can only be associated with one transcribed note n_j^{TR}. In our evaluation framework, we report Precision, Recall and F-measure:

$$\text{COnP}_{\text{Precision}}, \text{COnP}_{\text{Recall}} \text{ and } \text{COnP}_{\text{F-measure}}.$$

• Correct Onset (COn): Additionally, we have included the evaluation criteria used in MIREX *Audio Onset Detection* task. In this case, a note n_i^{GT} is assumed to be correctly transcribed into the note n_j^{TR} if it has correct onset. In addition, one ground truth note n_i^{GT} can only be associated with one transcribed note n_j^{TR}. In our evaluation framework, we report Precision, Recall and F-measure:

$$\text{COnPOff}_{\text{Precision}}, \text{COnPOff}_{\text{Recall}} \text{ and } \text{COnPOff}_{\text{F-measure}}.$$

3.2.4 Incorrect notes with one single error

In addition, we have included some novel evaluation measures to identify the notes that are close to be correctly transcribed, but they fail in one single aspect. These measures are useful to identify specific weaknesses of a given singing transcriber. The proposed categories are:

• Only-Bad-Onset (OBOn): A ground-truth note n_i^{GT} is labelled as OBOn if it has been transcribed into a note n_j^{TR} with correct pitch and offset, but wrong onset. In order to detect them, firstly we find all ground-truth notes with correct pitch and offset, taking into account that one ground-truth note can only be associated with one transcribed note. Then, we remove all notes previously tagged as COnPOff (Section 3.2.3). The reported measure is the rate of OBOn notes in the ground-truth:

$$\text{OBOn}_{\text{RateGT}}$$

• Only-Bad-Pitch (OBP): A ground-truth note n_i^{GT} is labelled as OBP if it has been transcribed into a note n_j^{TR} with correct onset and offset, but wrong pitch. In order to detect them, firstly we find all ground-truth notes with correct onset and offset, taking into account that one ground-truth note can only be associated with one transcribed note. Then, we remove all notes previously tagged as COnPOff (Section 3.2.3). The reported measure is the rate of OBP notes in the ground-truth:

$$\text{OBP}_{\text{RateGT}}$$

• Only-Bad-Offset (OBOff): A ground-truth note n_i^{GT} is labelled as OBOn if it has been transcribed into a note n_j^{TR} with correct pitch and onset, but wrong offset. In order to detect them, firstly we find all ground-truth notes with correct pitch and onset, taking into account that one ground-truth note can only be associated with one transcribed note. Then, we remove all notes previously tagged as COnPOff (Section 3.2.3). The reported measure is the rate of OBOff notes in the ground-truth:

$$\text{OBOff}_{\text{RateGT}}$$

3.2.5 Incorrect notes with segmentation errors

Segmentation errors refer to the case in which sung notes are incorrectly split or merged during the transcription. Depending on the final application, certain types of segmentation errors may not be important (e.g. frame-based systems for query-by-humming are not affected by splits), but they can lead to problems in many other situations. Therefore, we have defined two evaluation measures which are informative about the segmentation errors made by the singing transcriber.

• Split (S): A split note is a ground truth note n_i^{GT} that is incorrectly segmented into different consecutive notes $n_{j_1}^{TR}$, $n_{j_2}^{TR} \cdots n_{j_n}^{TR}$. Two requirements are needed in a split: (1) the set of transcribed notes $n_{j_1}^{TR}, n_{j_2}^{TR}, \ldots n_{j_n}^{TR}$ must overlap at least the 40% of n_i^{GT} in time (pitch is ignored), and (2) n_i^{GT} must overlap at least the 40% of every note $n_{j_1}^{TR}, n_{j_2}^{TR}, \ldots n_{j_n}^{TR}$ in time (again, pitch is ignored). These requirements are needed to ensure a consistent relationship between ground truth and transcribed notes. The specific reported measures are:

$$\text{S}_{\text{RateGT}} \text{ and } \text{S}_{\text{Ratio}}$$

Note that in this case $\text{S}_{\text{Ratio}} > 1$.

• Merged (M): A set of consecutive ground-truth notes $n_{i_1}^{GT}$, $n_{i_2}^{GT}$, $\cdots$ $n_{i_n}^{GT}$ are considered to be merged if they all are transcribed into the same note n_j^{TR}. This is the complementary case of split. Again, two requirements must be true to consider a group of merged notes: (1) the set of ground truth notes $n_{i_1}^{GT}$,$n_{i_2}^{GT}$, $\ldots$ $n_{i_n}^{GT}$ must overlap the 40% of n_j^{TR} in time (pitch is ignored), and (2) n_j^{TR} must overlap the 40% of every note $n_{i_1}^{GT}$,$n_{i_2}^{GT}$, $\ldots$ $n_{i_n}^{GT}$ in time (again, pitch is ignored). The specific reported measures are:

$$\text{M}_{\text{RateGT}} \text{ and } \text{M}_{\text{Ratio}}$$

Note that in this case $\text{M}_{\text{Ratio}} < 1$.

3.2.6 Incorrect notes with voicing errors

Voicing errors happen when an unvoiced sound produces a false transcribed note (spurious note), or when a sung note is not transcribed at all (non-detected note). This situation is commonly associated to a bad performance of the voicing stage within the singing transcriber. We have defined two categories:

• Spurious notes (PU): A spurious note is a transcribed note n_j^{TR} that does not overlap at all (neither in time nor in pitch) any note in the ground truth. The associated reported measure is:

$$\text{PU}_{\text{RateTR}}$$

• Non-detected notes (ND): A ground-truth note n_i^{GT} is non-detected if it does not overlap at all (neither in time nor in pitch) any transcribed note. The associated reported measure is:

$$\text{ND}_{\text{RateGT}}$$

3.3 Proposed Matlab toolbox

The presented evaluation measures have been implemented in a freely available Matlab toolbox [4], which consists of a set of functions and structures, as well as a graphical user interface to visually analyse the performance of the evaluated singing transcriber.

The main function of our toolbox is `evaluation.m`, which receives the ground-truth and the transcription of an audio clip as inputs, and it outputs the results of all the evaluation measures. In addition, we have included a function called `listnotes.m`, which receives as inputs the ground-truth, the transcription and the category **X** to be listed, and it outputs a list (in a two-columns format: onset time-offset time) of all the notes in the ground-truth tagged as **X** category. This information is useful to isolate the problematic audio excerpts for further analysis.

Finally, we have implemented a graphical user interface, where the ground-truth and the transcription of a given audio clip can be compared using a piano-roll representation. This interface also allows the user to highlight notes tagged as **X** (e.g. COnPOff, S, etc.).

4. PRACTICAL USE OF THE PROPOSED TOOLBOX

In this section, we describe a practical case of use in which the presented evaluation framework has been used to perform an improved comparative study of several state-of-the-art singing transcribers (presented in Section 4.1). In addition, a simple, easily reproducible baseline approach has been included in this comparative study. Finally, we show and discuss the obtained results.

4.1 Compared algorithms

We have compared three state-of-the-art algorithms for singing transcription:

Method (a): Gómez & Bonada (2013) [3]. It 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. For the experiment, we have used a binary provided by the authors of the algorithm.

Method (b): Ryynänen (2008) [13]. We have used the method for automatic transcription of melody, bass line and chords in polyphonic music published by Ryynänen in 2008 [13], although we only focus on melody transcription. It is the last evolution of the original HMM-based monophonic singing transcriber [9]. For the experiment, we have used a binary provided by the authors of the algorithm.

Method (c): Melotranscript [4] (based on Mulder 2004 [7]). It is the commercial version derived from the research carried out by Mulder et al. [7]. It is based on an auditory model. For the experiment, we have used the demo version available in SampleSumo website [3].

4.2 Baseline algorithm

According to [8], the simplest possible segmentation consists of simply rounding a rough pitch estimate to the closest MIDI note n_i and taking all pitch changes as note boundaries. The proposed baseline method is based on such idea, and it uses Yin [14] to extract the F0 and aperiodicity at frame-level. A frame is classified as unvoiced if its aperiodicity is under < 0.4. Finally, all notes shorter than 100ms are discarded.

4.3 Results & discussion

In Figure 2 we show the results of our comparative analysis. Regarding the F-measure of correct notes (COnPOff, COnP and COn), methods (a) and (c) attains similar values, whereas method (b) performs slightly worse. In addition, it seems that method (a) is slightly superior to method (c) for onset detection, but method (c) is superior when pitch and offset values must be also estimated. In all cases, the baseline is clearly worse than the rest of methods.

In addition, we observed that the rate of notes with incorrect onset (OBOn) is equally high (20%) in all methods. After analysing the specific recordings, we concluded that onset detection within a range of ± 50ms is very restrictive in the case of singing voice with lyrics, since many onsets are not clear even for an expert musician (as proved during the ground-truth building). Moreover, we also observed that all methods, and especially method (a), have problems with pitch bendings at the beginning of the notes, since they tend to split them.

Regarding the segmentation and voicing errors, we realised that method (a) tends to split notes, whereas method (b) tends to merge notes. This information, easily provided by our evaluation framework, may be useful to improve specific weaknesses of the algorithms during the development stage. Finally, we also realised that method (b) is worse than method (a) and (c) in terms of voicing.

To sum up, method (c) seems to be the best one in most measures, mainly due to a better performance in segmentation and voicing. However, method (a) is very appropriate for onset detection. Finally, although method (b) works clearly better than the baseline, has a poor performance due to errors in segmentation (mainly merged notes) and voicing (mainly spurious).

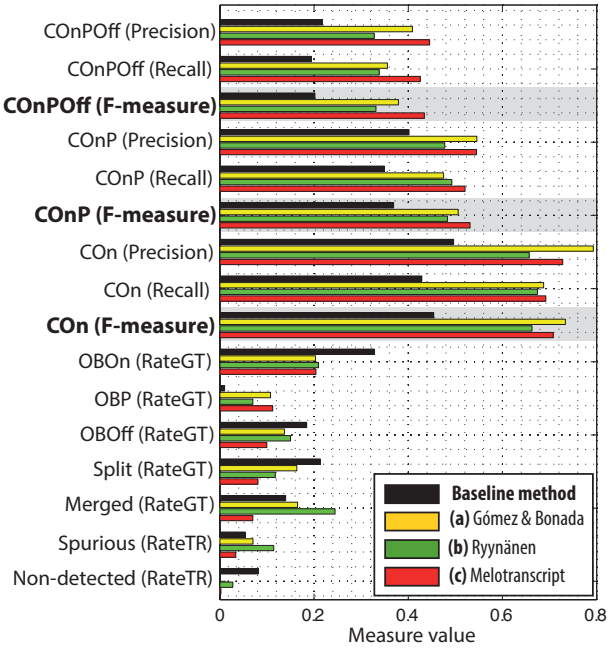

Figure 2. Comparison in detail of several state-of-the-art singing transcription systems using the presented evaluation framework.

5. CONCLUSIONS

In this paper, we have presented an evaluation framework for singing transcription. It consists of a cross-annotated dataset of 1154 seconds and a novel set of evaluation measures, able to report the type of errors made by the system. Both the dataset, and a Matlab toolbox including the presented evaluation measures, are freely available [4]. In order to show the utility of the work presented in this paper, we have performed an detailed comparative study of three state-of-the-art singing transcribers plus a baseline method, leading to relevant information about the performance of each method. In the future, we plan to expand our evaluation dataset in order to make it comparable to other datasets [7] used in MIREX (e.g. MIR-1K or MIR-QBSH).

6. ACKNOWLEDGEMENTS

This work has been funded by the Ministerio de Economía y Competitividad of the Spanish Government under Project No. TIN2013-47276-C6-2-R and by the Junta de Andalucía under Project No. P11-TIC-7154. The work has been done at Universidad de Málaga. Campus de Excelencia Internacional Andalucía Tech.

7. REFERENCES

[1] M. Ryynänen, "Singing transcription," in *Signal Processing Methods for Music Transcription* (A. Klapuri and M. Davy, eds.), pp. 361–390, Springer Science + Business Media LLC, 2006.

[2] E. Molina, I. Barbancho, E. Gómez, A. M. Barbancho, and L. J. Tardón, "Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, pp. 744–748, 2013.

[3] E. Gómez and J. Bonada, "Towards computer-assisted flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to a cappella singing," *Computer Music Journal*, vol. 37, no. 2, pp. 73–90, 2013.

[4] R. J. McNab, L. A. Smith, and I. H. Witten, "Signal Processing for Melody Transcription," *Proceedings of the 19th Australasian Computer Science Conference*, vol. 18, no. 4, pp. 301–307, 1996.

[5] G. Haus and E. Pollastri, "An audio front end for query-by-humming systems," in *Proceedings of the 2nd International Symposium on Music Information Retrieval ISMIR*, pp. 65–72, sn, 2001.

[6] L. P. Clarisse, J. P. Martens, M. Lesaffre, B. D. Baets, H. D. Meyer, and M. Leman, "An Auditory Model Based Transcriber of Singing Sequences," in *Proceedings of the 3rd International Conference on Music Information Retrieval ISMIR*, pp. 116–123, 2002.

[7] T. De Mulder, J.P. Martens, M. Lesaffre, M. Leman, B. De Baets, H. De Meyer, "Recent improvements of an auditory model based front-end for the transcription of vocal queries", , *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP 2004)*, Montreal, Quebec, Canada, May 17–21, Vol. IV, pp. 257–260, 2004.

[8] T. Viitaniemi, A. Klapuri, and A. Eronen, "A probabilistic model for the transcription of single-voice melodies," in *Proceedings of the 2003 Finnish Signal Processing Symposium FINSIG03*, pp. 59–63, 2003.

[9] M. Ryynänen and A. Klapuri, "Modelling of note events for singing transcription," in *Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing SAPA*, (Jeju, Korea), Oct. 2004.

[10] P. Kumar, M. Joshi, S. Hariharan, and P. Rao, "Sung Note Segmentation for a Query-by-Humming System". In *Intl Joint Conferences on Artificial Intelligence (IJCAI)*, 2007.

[11] W. Krige, T. Herbst, and T. Niesler, "Explicit transition modelling for automatic singing transcription," *Journal of New Music Research*, vol. 37, no. 4, pp. 311–324, 2008.

[12] J. Salamon, J. Serrá and E. Gómez, "Tonal Representations for Music Retrieval: From Version Identification to Query-by-Humming", *International Journal of Multimedia Information Retrieval, special issue on Hybrid Music Information Retrieval*, vol. 2, no. 1, pp. 45–58, 2013.

[13] M. P. Ryynänen and A. P. Klapuri, "Automatic Transcription of Melody, Bass Line, and Chords in Polyphonic Music," in *Computer Music Journal*, vol.32, no. 3, 2008.

[14] A. De Cheveigné and H. Kawahara: "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustic Society of America*, Vol. 111, No. 4, pp. 1917-1930, 2002.

[7] http://mirlab.org/dataSet/public/