AUTOMATIC SOLFÈGE ASSESSMENT

Rodrigo Schramm[1] **Helena de Souza Nunes**[2] **Cláudio Rosito Jung**[1]
[1] Institute of Informatics, Federal University of Rio Grande do Sul, Brazil
[2] Department of Music, Federal University of Rio Grande do Sul of Music, Brazil
rodrigos@caef.ufrgs.br, helena@caef.ufrgs.br, crjung@inf.ufrgs.br

ABSTRACT

This paper presents a note-by-note approach for automatic solfège assessment. The proposed system uses melodic transcription techniques to extract the sung notes from the audio signal, and the sequence of melodic segments is subsequently processed by a two stage algorithm. On the first stage, an aggregation process is introduced to perform the temporal alignment between the transcribed melody and the music score (ground truth). This stage implicitly aggregates and links the best combination of the extracted melodic segments with the expected note in the ground truth. On the second stage, a statistical method is used to evaluate the accuracy of each detected sung note. The technique is implemented using a Bayesian classifier, which is trained using an audio dataset containing individual scores provided by a committee of expert listeners. These individual scores were measured at each musical note, regarding the pitch, onset, and offset accuracy. Experimental results indicate that the classification scheme is suitable to be used as an assessment tool, providing useful feedback to the student.

1. INTRODUCTION

The practice of solfège is used by beginner musicians to learn and improve the ability of the musical reading through the repeated singing of musical notes from a music score. In fact, this kind of exercise is a fundamental part of the music learning process. It guides the student to build its own musical perceptions by creating an internal image of the sound along the vocal emission of a note (or sequences of notes as intervals, scales and melodies). The ability to read the notes on a music score and at the same time to hear internally and to sing them *a prima vista* is here generically called solfège, and it is considered a prerequisite for performance and effective musical knowledge [13]. During the solfège, is crucial to have a constant feedback by an external expert, who should be responsible for detecting eventual mistakes and pinpoint the best way to fix them. Traditionally, the evaluation process of the solfège is conducted by a music teacher, inside of a classroom. Nowadays, with the spread of the internet, new educational methods bring up new possibilities to music education using the e-learning paradigm. In the case of large number of evaluations, which is a typical situation in distance learning courses, the labor of the teacher becomes exhaustive and tedious. Even in cases of traditional and presential music lessons, the judgment by an expert musician is not a trivial task, specially because the human discernment may be affected by subjective factors and fatigue [14, 20]. Thus, an automatic solfège assessment tool can be very helpful in this context.

Usually, solfège is evaluated by comparing the singing performance with the target music score (ground truth). In this case, the first part of this task have some similarities with automatic melodic transcription algorithms [14]. However, a set of similarity measures to correlate the user performance with the expert's (human) judgment is still needed. Although there are some papers that provide an overall score for a given solfège [10], it is not to our knowledge the existence of systems that perform a note-by-note analysis, which is very important in music teaching. In this paper we introduce a new note-by-note evaluation method based on the individual scores provided by human evaluators. More precisely, we introduce a Bayesian classifier which is applied to each musical note detection, working as an alternative to the correlation method based on the global judgment score used in [10]. The main difference here is the fact that the performance can be evaluated with a small granularity, at each musical note, but keeping the assessment correlated with the human judgment. Additionally, the Bayesian approach allows the mapping of the performance errors into a confidence measure. We also introduced a new temporal alignment method between the transcribed melody and the music score (ground truth) by using a clustering process. The grouping process was chosen in place of dynamic time warping (DTW) [12] approach because it is less sensible to error propagation and it does not have any monotonicity condition.

This paper is organized as follows: Section 2 presents an overview about the related techniques. Section 3 shows a detailed description of the audio database generation and the corresponding annotation process by the musicians experts. Section 4 describes the proposed method to automatic solfège assessment. Section 5 presents the results of our experiments and Section 6 draws the conclusion of this work.

2. RELATED WORKS

As far as we know, there is no method for solfège evaluation in a note-by-note scale. Therefore, this section will revise some papers that tackle related problems. For example, Jha and Rao [4] focused on the vowel quality of the singing voice. The authors use low-level features, including the spectrum envelope and pitch contour for singing evaluation. Their algorithm detects the onset of each vowel by searching for rapid changes in specific frequency bands that characterize the vowel formants, and then correlates each vowel with an articulatory space by a linear regression scheme. Miryala et al. [8] do not perform assessment directly, but their approach automatically identifies vocal expressions as voice glides and vibratos, which could be also used as a kind of singing evaluation.

The related problem of melodic transcription has been studied by several researchers. The common pipeline on the melody transcription techniques splits the process in low-level feature estimation, note segmentation and labeling, and post processing [3, 11]. For example, [19] implemented a melodic transcription algorithm by detecting a sequence of fundamental frequencies in a frame-wise fashion, which are subsequently converted into observation probabilities and used in a Hidden Markov Model (HMM). Ryynanen and Klapuri [17] implemented a similar approach, but extending the number of low-level features. Thus, besides the fundamental frequency estimates, they also mapped into probabilities distributions the features regarding voice/ unvoice, accent, and meter estimation. Frequently, a musicological model is also included in music transcription algorithms to improve the system accuracy, acting as a prior probability. The authors of [19] also incorporate a duration model, which maps probability density functions with the subdivisions and multiples unities of the beat time. Musicological models might be used to detect the tonality and the rhythmic structure of the musical performance, constraining the output options and consequently improving the accuracy [5]. Unfortunately, the musicological model cannot be directly used as *a priori* information on assessment tools since it is not possible to have any expectation about the student singing performance.

The work by Molina et al. [10], which explores the singing assessment regarding note-based melodic similarities, as well as the temporal alignment between the student performance and the target melody, has similar goals to ours. Despite the use of note-level similarity measures, the final evaluation is built using the global assessment scores from the human experts, who had placed a global score (between 1 and 10) for each singing performance on a previous training stage. The estimated correlations in [10] seem to be advantageous to extract a measure of quality in a global context. However, as a drawback, that approach discard local (note level) information from the experts' evaluation. In other words, it is not possible to precisely locate and quantify the note(s) responsible for a bad or good score from the singing performance. A small extension of this approach was presented in [6], including new audio spectral features. A recent work [9] shows a taxonomy of evaluation measures used in several automatic singing transcriptions algorithms. Most of the tabulated approaches have used evaluation measures for singing transcription algorithms based on note/frame-level error. There are also some strategies that use time warping alignment information between the ground truth and the transcribed melody [10]. Despite the variety and effort to build robust and comprehensive evaluations measures, these previous ideas cannot be directly used in the context of solfège assessment. In fact, the used definition of correct pitch/ onset/offset in [9] applies ranges of tolerance with fixed values, that may be a reliable procedure to compare distinct algorithms of melodic transcription. However, it may not agree with the human judgment perception in a solfège assessment context. Some authors [6, 10] tried to solve this issue connecting the expert analysis with the evaluation measures, but the final human evaluation carries out only a global interpretation, lacking in details at individual sung notes. In the next sections we present our dataset and proposed model for solfège assessment. This model aims to evaluate individual sung notes, giving a note-based feedback that makes a meaningful link with the human judgment by musician experts.

3. PROPOSED DATASET AND ANNOTATIONS

The proposed dataset consists of sequences of musical intervals in the chromatic scale. The audio recordings were done using seven adults, including trained (three) and untrained (four) singers ranging from 17 to 61 years old. These melodic sequences were recorded during four months, in mono format with a sample rate of $44100Hz$ and 16 bits quantization.

It was decided to support the singing process by a reference piano audio track, since a part of the group of singers was unable to read music scores. In this reference audio track, the intervals were played in sequence, but with gaps between them. Each singer filled these gaps repeating the previous heard melodic interval at the next beat time, and all recording sessions were synchronized by a metronome.

The singers were asked to choose and, if possible, to diversify the used phonemes. They were also asked to sing freely, but respecting the pitch, attack and duration of the previously indicated sounds, aiming to capture real examples of spontaneous everyday singing. Intentionally aiming to capture a higher variability of natural situations, the recordings were conducted in two distinct environments: a part of the examples was recorded in a studio, where the resulting audio records are clean; another part of the audio records was done in informal conditions, presenting some background noise and reverberation.

A total of 21 sessions were recorded, containing (twelve ascending intervals and twelve descendants intervals of the chromatic scale). Each singer performed the melodic intervals in three distinct tempos: Adagio, 60 bpm; Andante Moderato, 90 bpm; and Allegro, 120 bpm. Along with the recordings, an annotation process was conducted by a committee of experts (five graduated musicians with more than ten years of experience in solfège assessment auditions) in order to label each sung note from the recorded dataset into two possible categories (correct and incorrect) regarding the pitch, onset and offset accuracy. Before each annotation section, the committee was advised to hear some random samples from the dataset. This warmup procedure was important because it helped to create an agreement among the experts, who shared some important characteristics and aspects of the recorded melodic intervals. As the dataset was broken in parts, this process was repeated in several days, until the whole set of audio records had been evaluated (in fact, the whole process for building the annotated dataset took several months).

For each sung note in the audio dataset, all the five evaluators casted a vote (correct or incorrect) regarding each analyzed parameter (pitch, onset and offset). As it will be explained in the next section, disagreement among the evaluators were kept and used to model our probabilistic classifier. Also, each note is assigned to a single label (correct or incorrect) regarding to each parameter, based on the majority of votes cast by the experts (i.e., at least 3 votes for the same label). Hence, some labels can be considered more reliable than others, based on the number of votes. For example, regarding the pitch, 15.38% of the samples received 3 votes in agreement, which means an expressive degree of doubt among the experts. The same analysis was made for the onset and offset parameters, and the percentage of notes with 3 votes (doubt) was 10.71% and 12.09%, respectively. The final annotated dataset contains 3276 labeled samples.

4. OUR MODEL

The proposed computational model for automatic solfège evaluation is structured in two main stages. The first stage performs the melodic transcription, using the pYIN algorithm [7] to extract the fundamental frequency from the audio signal. The pYIN algorithm is a modification of the smoothing procedure of the YIN technique [1], introducing a probabilistic variant that outputs multiple pitch candidates along with the associated probabilities. It also employs a Hidden Markov Model (HMM) based on [16] to perform the pitch tracking, providing an improvement in the accuracy of the standard YIN. The extracted frame-wise sequence f_0 is then segmented and labeled into segments of music notes using the hysteresis approach of [11]. After, in this stage, we introduced a new alignment procedure, where the transcribed sung segments are aligned with the music score. This procedure converts group of melodic segments into atomic unities (music notes) and allows a direct comparison (note against note) between the transcribed melody and the ground truth. In the second stage, a probabilistic classifier performs the note-based evaluation. The algorithm takes the generated sequence of notes from the previous stage and applies a Bayesian classifier to evaluate the accuracy of the parameters pitch, onset and offset. At this stage, a rejection procedure is also introduced to map the doubt from the categorization (correct or incorrect) given by the expert listeners. These stages are described next.

4.1 Melodic alignment

To evaluate the solfège performance, a comparison of the singing performance with the target music score (ground truth) is required. Thus, after obtaining the automatic melodic transcription (using [7] and [11]), it is still necessary to connect each transcribed note with its corresponding note in the music score.

The first challenge is the fact that the melodic transcription often generates groups of fragmented notes (segments), which should be mapped to only one element of the ground truth. Each melodic fragment is represented by f_{il}, where i is the segment index and l is the relative index of each frame within this segment. Additionally, as in [10], there is no assumption of synchronization by a metronome in our approach, so that the transcribed notes might be misaligned. In [10], an integrated dynamic time warping procedure (DTW) was employed to perform the time alignment in a frame-wise fashion. However, in some cases, the boundary condition of the DTW algorithm might propagate the accumulated matching error, which causes an undesirable alignment between the transcribed sequence and the ground truth.

Here, we propose a new alignment process that, at the same time, groups note fragments and also maps the resulting block with the correspondent music note in the ground truth. Despite being similar to the DTW approach, it does not propagate the cumulative error since it does not need to obey the boundary condition of the DTW algorithm. The joint grouping/alignment process was designed as a brute force algorithm that is implemented using a cost matrix C. For each note k in the ground truth, the algorithm computes the cumulative distance measure considering all possibilities of grouping of adjacent segments, starting at segment index i and stopping at segment index j. This algorithm is efficiently built with the support of a 3D data structure, as depicted in Figure 1a. Thus, for each possible combination (k,i,j), a dissimilarity measure is computed as

$$\begin{aligned} C(k,i,j) = \alpha_1 \Delta f(k,i,j) + \alpha_2 \Delta d(k,i,j) + \\ \alpha_3 \Delta s(k,i,j) + \alpha_4 \Delta e(k,i,j), \end{aligned} \tag{1}$$

where

$$\Delta f = |f_k^{gt} - median(f_{i,1} \dots f_{j,l_{max}})| \tag{2}$$

is the pitch distance between the ground truth note k and the median values of f_0 belonging to the range starting at first frame of the segment i and finishing at the last frame l_{max} of the segment j,

$$\Delta d = |D_k^{gt} - \sum_{m=i}^{j} D_m| \tag{3}$$

measures the duration difference (in seconds) between the note k in the ground truth (D_k^{gt}) and the group formed from segment i to j in the transcribed melody (D_i is the duration of segment i),

$$\Delta s = |S_k^{gt} - S_i| \tag{4}$$

accounts for the delay or advance (in seconds) of the onset of the first segment of the selected group and the ground

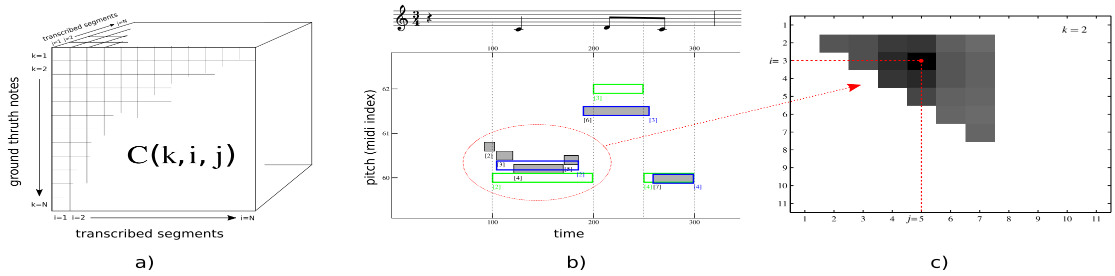


Figure 1: (a) 3D structure used to compute the similarities between the transcribed melodic segments and the music score. (b) Grouping process of several segments (gray) into one music note (blue). (c) The best grouping for the note k in the ground truth is found by the indexes i (first element) and j (last element), which minimize the function $C(k,i,j)$.

truth note k, and analogously

$$\Delta e = |E_k^{gt} - E_j| \tag{5}$$

accounts for the delay or advance of the offset (in seconds). The coefficients α_i are weights to balance the individual contribution of each measure, and our experiments show that $\alpha_1 = 1.0$, $\alpha_2 = 2.0$, $\alpha_3 = 2.0$, $\alpha_4 = 2.0$ is a good combination.

The grouping process and its mapping to the ground truth sequence is achieved by a function

$$\upsilon(k) = (i_k, j_k) = \underset{i,j}{\operatorname{argmin}} C(k,i,j), \tag{6}$$

so that each note k is mapped to the group of segments from indices i (first segment) to j (last segment), obtaining the final and consolidated transcribed note.

The computational complexity of the alignment process in the worst case is $\mathcal{O}(MN^2)$, where M is the number of music notes in the ground truth and N is the number of melodic segments. However, the inclusion of components Δs and Δe in Eq. (1) makes the magnitude of the dissimilarity measure to grow fast when the group of segments is far from the expected time position. As a consequence, it is possible to interrupt the brute force search loop in a few iterations by limiting the value of $C(k,i,j)$. Furthermore, the window of evaluation containing the melodic segments can be restricted to begin closer to the target note. This process will also decrease the computational cost and also avoid eventual local minimum issues in Eq. (6). Figure 1b illustrates one example of the grouping and alignment process, in which six segments are mapped into three notes.

4.2 Note-based evaluation

After the alignment achieved by the melodic transcription, the system performs the note-based assessment. Distinct probability density functions are modeled to represent the correct and incorrect sung notes, regarding individually to the pitch (Δf, in midi scale), onset (Δs, in seconds) and offset (Δe, in seconds) deviations. For each sung note, a Bayesian classifier assigns the parameters pitch, onset and offset into correct φ or incorrect $\overline{\varphi}$ categories. Next, the Bayesian classification process will be explained, focusing on the Δf (pitch) parameter. However, it is worth noting that the classification process is also individually applied to Δs and Δe in an analogous way.

Figure 2a shows the histograms of the pitch deviations for correct and incorrect categories based on the expert's evaluation, denoted by $\varphi_{\Delta f}$ and $\overline{\varphi}_{\Delta f}$, respectively. As it can be observed, the histogram of $\varphi_{\Delta f}$ presents a sharp peak close the origin (related to low pitch errors), as expected. Nevertheless, the two categories present considerable overlap, corroborating the discrepancies in the accuracy evaluation by experts when for intermediate errors in the pitch. In fact, since we had used the individual ratings of each note from all evaluators to build de histograms, the pitch deviation Δf related to a note that received conflicting labels among the evaluators contributes both for the histograms of $\varphi_{\Delta f}$ and $\overline{\varphi}_{\Delta f}$.

A conditional probability density function is then estimated from the distributions of Δf for each class $r \in \{\varphi_{\Delta f}, \overline{\varphi}_{\Delta f}\}$, so that a posterior probability (that can be considered a measure of confidence) can be easily obtained. Among several existing parametric probability density functions (PDFs) for modeling positive random variables, the Gamma distribution was chosen because it has been successfully used to model similar problems [18], which have similar characteristics to our data, such as single mode and frequently skewed shape. The *gamma* PDF, parameterized by the two positive parameters shape α_r and scale θ_r, is given by:

$$p(\Delta f|r) \sim Ga(\Delta f; \alpha_r, \theta_r) = \frac{\Delta f^{\alpha_r - 1} e^{\frac{-\delta_r}{\theta_r}}}{\Gamma(\alpha_r)\theta_r^{\alpha_r}}, \tag{7}$$

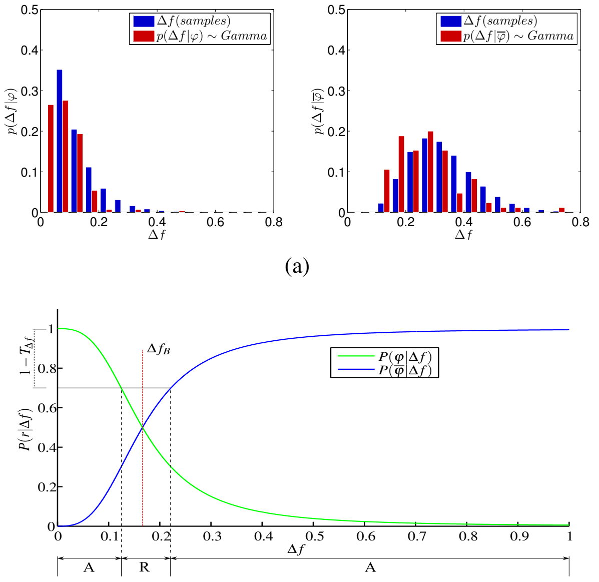
where Γ is the *gamma* function.

The shape (α_r) and scale (θ_r) parameters for each class $r \in \{\varphi, \overline{\varphi}\}$ were estimated using a maximum likelihood approach [15]. Given the PDFs $p(\Delta f|\varphi)$ and $p(\Delta f|\overline{\varphi})$, we can estimate the *posterior* probability of the pitch of a correct/incorrect sung note by using the Bayes rule [2]:

$$p(r|\Delta f) = \frac{p(\Delta f|r)P(r)}{p(\Delta f)}, \tag{8}$$

where $p(\Delta f) = p(\Delta f|\varphi)P(\varphi) + p(\Delta f|\overline{\varphi})P(\overline{\varphi})$ is the overall distribution of Δf, and the *prior* probabilities $P(\varphi)$ and $P(\overline{\varphi})$ are defined as equiprobable.

Figure 2b illustrates the decision boundary for $\varphi_{\Delta f}$ and $\overline{\varphi}_{\Delta f}$ as a red vertical dashed line, and it can be observed



(a)

(b)

Figure 2: (a) Histogram of Δf for classes φ and $\overline{\varphi}$ along with fitted Gamma PDFs. (b) Posterior probabilities, along with acceptance and rejection regions.

that there is a "fuzzy" decision boundary around it. In this region, there is considerable overlap between $p(\Delta f|\varphi)$ and $p(\Delta f|\overline{\varphi})$, causing the winning posterior probability to be just a little above 0.5. Since this overlap region is caused in part by conflicting labels from the evaluators, an appropriate option is to reject samples that fall inside this fuzzy region. As in [18], the errors (or misclassifications) are converted into rejects using the Bayes rejection rule for the minimum error [21]. The rejection rule splits the sample space into an acceptance region A and a rejection region R, that is given by:

$$R(T_{\Delta f}) = \{\Delta f | 1 - \max_r p(r|\Delta f) > T_{\Delta f}\}, \tag{9}$$

$$A(T_{\Delta f}) = \{\Delta f | 1 - \max_r p(r|\Delta f) \leq T_{\Delta f}\}, \tag{10}$$

where the threshold $T_{\Delta f}$ balances the tradeoff between the number of rejected samples and the error rate $e(T_{\Delta f})$, given by:

$$e(T_{\Delta f}) = \sum_{\Delta f \in A(T_{\Delta f})} \left(1 - \max_r p(r|\Delta f)\right) p(\Delta f). \tag{11}$$

The choice of the threshold $T_{\Delta f} = 0.33$ was determined from a set of experiments where the classification accuracy and the number of rejections were taken into account (more details about this choice are presented in section 5). The posterior probabilities and the boundaries between regions A and R generated by this threshold are shown in Figure 2b.

Thus, regarding the pitch accuracy and using the Bayesian classifier given by Eq. (8) in combination with the rejection procedure provided by Eqs. (9) and (10), each sung note is classified into three possible classes: correct, incorrect, or undetermined (reject). When a classification is done (correct or incorrect), the corresponding probability measure is also used to provide a meaningful feedback of confidence to the user. The whole note-based evaluation process is also done independently for the onset and offset note accuracy. This means that, for each sung note, the system output gives individual class labels and confidence measures for pitch, onset and offset.

5. EXPERIMENTAL RESULTS

Aiming to extract an objective evaluation of the proposed solfège assessment system, a set of experiments were conducted using the annotated audio dataset described in Section 3. From the audio recordings, we extracted the melodic transcriptions, which were subsequently aligned with the ground truth, as described in Section 4.1. The pitch, onset and offset deviations (Δf, Δs and Δe) were computed from the comparison between the ground truth and the aligned melodies, and a subset of the samples was used to estimate the parameters required in the corresponding Gamma PDFs. The remaining samples were reserved to test the model.

For the validation scheme, we used a 10-fold cross-validation scheme, in which the dataset is split randomly into ten equal parts. For each round of the cross-validation, 9 folds are used to train the probabilistic model and the remaining fold is used to validate the Bayesian classifier described in Section 4.2. In our experiments, we used the Bayesian classifier with and without the rejection rule. In both situations, the system classifies each parameter (pitch, onset, offset) of each sung note in two possibles categories: correct or incorrect (when the rejection rule was applied, some notes were kept unclassified).

Table 1 shows the confusion matrices generated by the Bayesian classifiers for the pitch, onset and offset without the rejection rule, and the accuracy is over 90% for the three analyzed parameters. Also, the system tends to produce more false negatives (i.e., mark as incorrect a correctly sung note) then false positives, particularly for the offset parameters, being a "rigid" evaluator. The misclassification errors are caused by two main reasons: first, a possible bad melodic transcription and/or bad alignment between the sung fragments and the ground truth can introduce errors on the similarities measures; second, the disagreement between the human evaluators generated an inherently fuzzy region near to the decision boundary. In fact, as noted in Section 3, 10 to 15% of the notes presented strong disagreement among the evaluators, so that the ground truth label may not be reliable.

The rejection rule provided by Eq. 9 avoids the classification of samples that potentially fall inside this fuzzy region. The effect of varying the rejection thresholds in the percentage of accepted samples and also the accuracy for the pitch, onset and offset analysis is shown in Figure 3. As expected, lower thresholds decrease the number of accepted samples and increases the accuracy rate. Although the definition of an optimal value for the threshold is difficult, the accuracy should be as maximum as possible while the number of rejected samples should be minimal. As the focus of this work is on music education, we believe it is

Target Class	Output Class: $\varphi_{\Delta f}$	$\overline{\varphi}_{\Delta f}$	
$\varphi_{\Delta f}$	**88.99%**	11.01%	
$\overline{\varphi}_{\Delta f}$	7.27%	**92.73%**	
			90.86%

(a) Pitch evaluation

Target Class	Output Class: $\varphi_{\Delta S}$	$\overline{\varphi}_{\Delta S}$	
$\varphi_{\Delta S}$	**89.17%**	10.83%	
$\overline{\varphi}_{\Delta S}$	8.74%	**91.26%**	
			90.22%

(b) Onset evaluation

Target Class	Output Class: $\varphi_{\Delta E}$	$\overline{\varphi}_{\Delta E}$	
$\varphi_{\Delta E}$	**84.71%**	15.29%	
$\overline{\varphi}_{\Delta E}$	2.54%	**97.46%**	
			91.08%

(c) Offset evaluation

Table 1: Evaluation of the proposed approach using 10-Folds cross validation without the Bayesian rejection rule.

Target Class	Output Class: $\varphi_{\Delta f}$	$\overline{\varphi}_{\Delta f}$	
$\varphi_{\Delta f}$	**94.45%**	5.55%	
$\overline{\varphi}_{\Delta f}$	2.54%	**97.46%**	
			95.96%

(a) Pitch evaluation: $T_{\Delta f} = 0.33$

Target Class	Output Class: $\varphi_{\Delta S}$	$\overline{\varphi}_{\Delta S}$	
$\varphi_{\Delta S}$	**94.17%**	5.83%	
$\overline{\varphi}_{\Delta S}$	7.34%	**92.66%**	
			93.42%

(b) Onset evaluation: $T_{\Delta S} = 0.31$

Target Class	Output Class: $\varphi_{\Delta E}$	$\overline{\varphi}_{\Delta E}$	
$\varphi_{\Delta E}$	**91.64%**	8.36%	
$\overline{\varphi}_{\Delta E}$	2.54%	**97.46%**	
			94.55%

(c) Offset evaluation: $T_{\Delta E} = 0.39$

Table 2: Evaluation of the proposed approach using 10-Folds cross validation with the Bayesian rejection rule. The system can answer in 90% of the times, increasing the final accuracy in almost 4%.

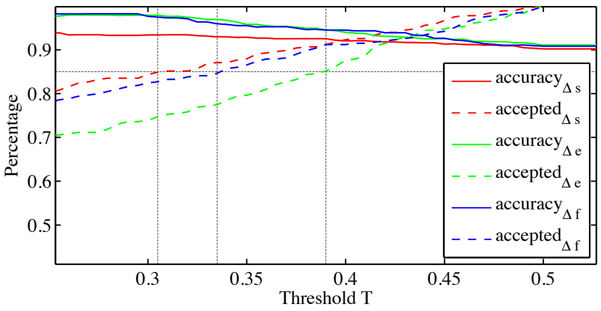


Figure 3: Comparative of the accuracy versus the number of non-rejected samples. Solid lines show the accuracy evolution, which are affected by the thresholds $T_{\Delta f}$ (pitch), $T_{\Delta s}$ (onset), and $T_{\Delta e}$ (offset).

preferred to not have an answer than to provide an incorrect feedback. Based on this assumption, and also considering that the percentage of samples with doubt from the expert evaluation is over 10%, we decided to set all thresholds to reject 15% of the samples in average.

Table 2 shows the accuracy evaluation for the 10-fold experiment using the Bayesian classifier with the rejection rule, in which the rejection thresholds $T_{\Delta f}$ (pitch), $T_{\Delta S}$ (onset) and $T_{\Delta E}$ (offset) were set so that 15% of the samples are rejected, matching approximately the percentage of samples with dubious labels. As it can be observed, the overall accuracies for all analyzed parameters increased in 3-5% when compared to the option without rejection, reaching up to almost 96% accuracy. Also, the number of false negatives was greatly reduced, particularly for the offset evaluation. This fact indicates that when in doubt, the evaluators tend to label a note as correct rather than incorrect. Furthermore, 32–35% of the rejected samples received 3 agreeing votes by the experts, which means that our system is removing more than twice of the samples related to the experts' doubt when compared with the whole dataset.

6. CONCLUSION

This paper presented a note-by-note approach for automatic solfège assessment focused on musical education, in which each sung note is evaluated considering the human evaluation perception in small scale, focused on the parameters of pitch, onset and offset at a specific part of the solfege practice. The proposed system uses melodic transcription techniques to extract the sung notes from the audio signal, and the sequence of melodic segments is subsequently processed by a two stage algorithm. In the first stage, an aggregation process was introduced to perform the temporal alignment between the transcribed melody and the music score (ground truth). This stage implicitly aggregates and links the best combination of the extracted melodic segments with the expected notes in the ground truth. The proposed alignment process does not impose the DTW boundary condition between the two sequences, avoiding the propagation of the accumulated matching error. In the second stage, a Bayesian classifier is used to evaluate the accuracy of each detected sung note. This statistical model was trained using a combination of the extracted measures (Δf, Δs, and Δe) with the individual scores provided by a committee of expert listeners.

Experimental results indicate that the classification scheme achieved accuracy rates in the range 90–91% without using the rejection rule (i.e., feedback for all evaluated notes), and 93–96% using the Bayesian rejection procedure (for the chosen thresholds, our tool is able to give feedback in 85% of the trials in average). Besides the classification label (correct, incorrect or undefined), the system also provides probability measure, which helps to indicate how likely correct or incorrect was the performance of the sung note. As future work, new research is planned to integrate new audio features, as well as the usage of lyrics analysis, to improve the segmentation and alignment on the first stage of this approach.

Acknowledgements: Thanks to CAPES Foundation, Ministry of Education of Brazil, scholarship BEX-2106/13-2.

7. REFERENCES

[1] Alain de Cheveigné and Hideki Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[2] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, pages 24–27, 2001.

[3] Emilia 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*, 37:73–90, 2013.

[4] Mayank Vibhuti Jha and Preeti Rao. Assessing vowel quality for singing evaluation. In *Proceedings of the National Conference on Communications (NCC)*, pages 1–5, Kharagpur, India, Feb 2012.

[5] Anssi Klapuri and Manuel Davy. *Signal Processing Methods for Music Transcription*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, pages 361–390, 2006.

[6] Chang-Hung Lin, Yuan-Shan Lee, Ming-Yen Chen, and Jia-Ching Wang. Automatic singing evaluating system based on acoustic features and rhythm. In *Proceedings of the IEEE International Conference on Orange Technologies (ICOT), 2014*, pages 165–168, Sept 2014.

[7] Matthias Mauch and Simon Dixon. pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, pages 659–663, May 2014.

[8] Sai Sumanth Miryala, Ranjita Bhagwan, Monojit Choudhury, and Kalika Bali. Automatically identifying vocal expressions for music transcription. In *Proceedings of the 14th International Society of Music Information Retrieval, ISMIR 2013, Curitiba, Brazil, November 4-8*, pages 239–244, 2013.

[9] Emilio Molina, Ana M. Barbancho, Lorenzo J. Tardón, and Isabel Barbancho. Evaluation framework for automatic singing transcription. In *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31*, pages 567–572, 2014.

[10] Emilio Molina, Isabel Barbancho, Emilia Gómez, Ana Maria Barbancho, and Lorenzo J. Tardón. Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP, 2013)*, pages 744–748, May 2013.

[11] Emilio Molina, Lorenzo J. Tardón, Ana M. Barbancho, and Isabel Barbancho. Sipth: Singing transcription based on hysteresis defined on the pitch-time curve. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(2):252–263, Feb 2015.

[12] Meinard Müller. *Information Retrieval for Music and Motion*. Springer Verlag, pages 69–74, 2007.

[13] Dorothy Payne. Essential skills, part 1 of 4: Essential skills for promoting a lifelong love of music and music making. *American Music Teacher*, February–March 2005.

[14] Graham E. Poliner, Daniel P. W. Ellis, A. F. Ehmann, Emilia Gómez, S. Streich, and Beesuan Ong. Melody transcription from music audio: Approaches and evaluation. *IEEE Transactions on Audio, Speech and Language Processing*, 15(4):1247–1256, 2007.

[15] Kandethody M. Ramachandran and Cris P. Tsokos. *Mathematical Statistics with Applications*. Elsevier Academic Press, 2009.

[16] Matti Ryynänen. Probabilistic modelling of note events in the transcription of monophonic melodies. Master's thesis, Tampere University of Technology, March 2004.

[17] Matti Ryynänen and Anssi Klapuri. Modelling of note events for singing transcription. In *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio*, Jeju, Korea, October 2004.

[18] Rodrigo Schramm, Cláudio Rosito Jung, and Eduardo Reck Miranda. Dynamic time warping for music conducting gestures evaluation. *Multimedia, IEEE Transactions on*, 17(2):243–255, Feb 2015.

[19] Timo Viitaniemi, Anssi Klapuri, and Antti Eronen. A probabilistic model for the transcription of single-voice melodies. In *Tampere University of Technology*, pages 59–63, 2003.

[20] Joel Wapnick and Elizabeth Ekholm. Expert consensus in solo voice performance evaluation. *Journal of Voice*, 11(4):429 – 436, 1997.

[21] Andrew R. Webb. *Statistical Pattern Recognition*. Wiley, Chichester,UK, 3 edition, pages 8–17, 2011.