Automatic Quality Assessment of Vocal and Instrumental Performances of Ninth-grade and Tenth-grade Pupils

Jakob Abeßer[1], Johannes Hasselhorn[2], Christian Dittmar[1], Andreas Lehmann[2], and Sascha Grollmisch[1⋆]

[1] Semantic Music Technologies, Fraunhofer IDMT, Germany
[2] Hochschule für Musik, Würzburg, Germany
jakob.abesser@idmt.fraunhofer.de

Abstract. We present our approach to the assessment of practical music competency among high-school age pupils. First, we describe our setup that ensures controlled and reproducible conditions for simultaneous performance assessments with class sizes of up to 25 pupils. Second, we outline the signal processing and machine learning tasks involved in the automatic modeling of experts ratings for music proficiency. The evaluation methodology and results are discussed in detail.

Keywords: Automatic music performance assessment, melody transcription, transcription-based audio features, intonation, group testing

1 Introduction

Music making is an integral part of music education in schools. It also forms the backbone of cultural participation in adulthood. In different fields of research, such as music education and music therapy, the assessment of music performance (abilities) may be of interest. Music making is traditionally evaluated on an individual basis, resulting in impractical testing durations and procedures. Rarely is it evaluated in a group environment, much less in large-scale assessments (e.g. [14]). This problem could be remedied by simultaneous assessment of all group members. Furthermore, judging performances can be a time-consuming task. For example, a particular music teacher assessing five school classes, each consisting of 25 pupils performing for 5 minutes, would have to listen to over 10 hours of recorded material. Therefore, both a simultaneous recording of all pupils and an automatic evaluation tool would be desirable when performing large-scale evaluation experiments.

2 Goals

Our goal was to measure the music making skills of pupils in German grade school courses within the framework of competency modeling [7]. More precisely,

⋆ This work was funded by the Deutsche Forschungsgemeinschaft (DFG).

we wanted to record vocal and instrumental music performances and develop a system for the automatic assessment of those recorded performances.

Using an automatic melody transcription algorithm and annotations of the performance quality by music experts, we applied a feature-selection algorithm to identify the most discriminatory audio features that could be used to train a statistical classification model of the experts' ratings.

3 Challenges

Vocal timbres possess a high variability and depend on age, gender, and degree of musical training. Findings in the literature and music experts opinions show low agreement in what constitutes desirable vocal characteristics. Here, we focus on vocal performance of pupils whose voices are representative of the largely untrained population. In the recording sessions, we often encountered timid pupils who, rather than singing, recited the lyrics of a given song. This led to erroneous automatic melody transcription results.

4 Previous Approaches

Mitchell and Kenny [11] showed that the assessment of vocal performances is challenging and difficult. Not even professional singers always agree on attributes that describe voice quality. Salvator provided an extensive overview of singing voice assessment methods and instruments in [15].

Hornbach and Taggart developed a 5-point assessment rubric to assess elementary-age children [6]. The authors reported satisfactory interjudge reliabilities (r=0.76 to r=0.97). However, this rubric only captures an overall impression of quality without specifying underlying factors (e.g., breathing, phrasing, tone etc.).

In [9] and [8], Larrouy-Maestri et al. analyzed a corpus of 166 vocal recordings of *Happy Birthday* by occasional singers. The authors automatically extracted the fundamental frequency curve and computed three features from the number of contour errors, the interval deviations, and the number of tonality modulations. Expert judges annotated the vocal accuracy of each recording on a 9-point scale. A significant agreement (0.77) was found among the expert judges. In a multiple linear regression analysis, both the interval deviation and the number of tonality modulations contributed significantly to the regression model.

Molina et al. also tried to automatically evaluate vocal recordings in [12]. To measure performance quality, sung melodies were rated by the authors on a 10-point scale in comparison with a given reference melody using as criteria intonation, rhythm, and overall judgment. Two approaches were used for the automatized scoring: First, both fundamental frequency curves of the transcribed vocal recording and from the reference melody were temporally aligned using Dynamic Time Warping (DTW). Based on the deviation of the optimal DTW path in the similarity matrix, two similarity measures for intonation and rhythmic errors were derived. It is worth noting that the authors constrained the maximum

deviation of the optimal DTW path from the main diagonal. They assumed that the deviations between the sung and the reference melody were only moderate. However, in our work, we cannot make such an assumption. Second, the automatic melody transcription was compared to the reference melody using different note-level similarity measures. Six similarity measures were derived from pitch, interval, and onset deviations between both melodies. The authors found a very high inter-judgment reliability for all three ratings. Quadratic polynomial regression was used to model the expert ratings by means of the computed feature values. Based on their data set of 27 vocal recordings, the authors reported very high correlation coefficients around 0.98 and small RMSE values between 0.41 and 0.58.

5 Novel Approach

Figure 1 gives an overview of our approach to the automatic assessment of music performance. First, we recorded pupils who had to perform different vocal and instrumental tasks. The data acquisition procedure is explained in Section 5.1. Then, each recorded performance was annotated by multiple music experts as described in Section 5.2. As shown in Section 5.3 and 5.4, the automatic analysis consisted of two components—an automatic melody transcription stage and a feature extraction stage. Based on the extracted features and the experts performance ratings, the most discriminatory features were selected and used to train a classification model of the ratings as detailed in Section 6.2.

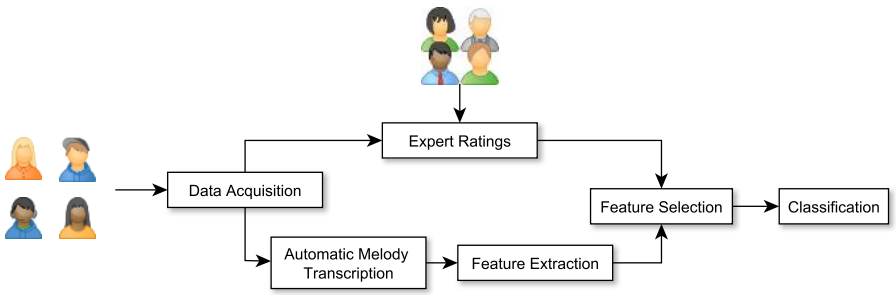


Fig. 1. Flowchart of the proposed framework.

5.1 Data Acquisition

In this paper, we focus on vocal and instrumental recordings. All recordings were conducted in German schools during class time. In each recording session, up to 15 pupils from ninth- or tenth-grade were recorded simultaneously. Each student had to perform multiple vocal and instrument recording tasks.

Vocal recordings The vocal recordings were undertaken using headphones for the playback of instructions and play-along with an attached headset microphone for the actual voice recording.

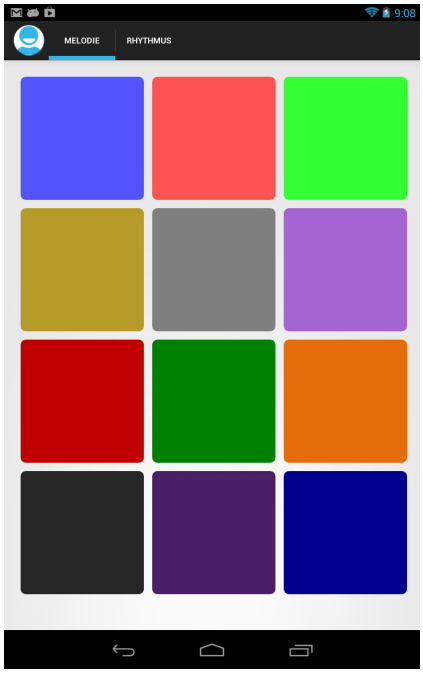


Fig. 2. Screenshot of the "Colored Music Grid" (CMG) app.

Instrument recordings We designed a new musical input device (henceforth called Colored Music Grid" (CMG) app), which is used for the musical instrument recording tasks. It compensates for the potential advantage that pupils with antecedent music instruction on keyboard instruments may have. The instrument is implemented as a music application, which can be used on tablet computers with a multi-touch surface (see screenshot in Figure 2).[3] The colored fields of the grid correspond to a diatonic scale (e.g., C-major). The playable notes span a pitch range of about one and a half octaves and are aligned like a numeric keypad. The same colors stand for the same pitch class; hues indicate pitch height (an octave apart) such as light and dark red for C1 and C2, respectively. During the instrument recording tasks, the students always heard an accompaniment track on their headphones. The students' task was to play the melody given by the score to the accompaniment track in a musically fitting

[3] In our data acquisition procedure, 7-inch tablet computers were used.

Original rating class	1	2	3	4	5	6
Rating class after mapping	1	2	2	3	3	4

Table 1. Mapping for the reduction of number of rating classes.

way. Here, the note pitch values were colored similarly as in the CMG app. An example task is shown in Figure 3.

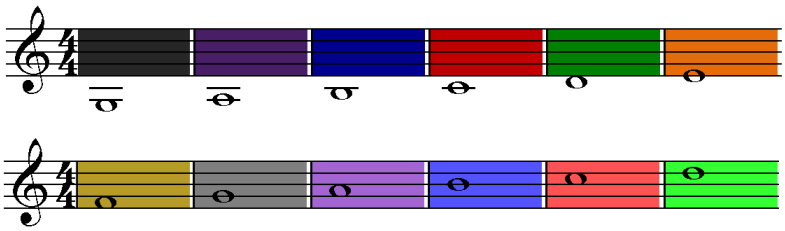


Fig. 3. Example instrument tasks for the CMG app.

5.2 Expert Ratings

The recorded performances were assessed using two new rating scales that are shown in Table 2. Both scales are based on the 5-point Hornbach and Taggert rubric in [6]. The ratings were performed by 3 music teachers and 7 to 10 music students who showed a high agreement in their annotations. The intra-class correlation (ICC) values ranged from .74 to .95 for the vocal recordings and .79 and .97 for the recordings with the CMG app. In order to facilitate the automatic performance assessment, we reduced the number of classes from 6 to 4 using the mapping shown in Table 1.

5.3 Automatic Melody Transcription

We use an existing algorithm [4] for extraction of the pre-dominant melody throughout the recordings. The algorithm is based on frame-wise computation of pitch saliency spectra as described in [3]. The underlying time-frequency transform is implemented as an efficient Multi-Resolution Fast Fourier Transform (MRFFT) on overlapping frames with a hop-size of approximately 6 ms. The MRFFT simultaneously enhances sinusoidal peaks in the lower frequency range and takes into account rapid changes in the upper frequency range (for instance those caused by vibrato) [2]. The evolution of the most salient pitch candidates is tracked over time by multiple auditory streams that later form tone objects.

Rating	Vocal rating scale	Instrumental rating scale
1	Child is a nearly or totally accurate singer	Child plays instrumental part nearly or entirely correctly (with a clear sense of musical meaning)
2	Child sings with some accuracy, beginning in the established key	Child plays instrumental part nearly or entirely correctly (with little sense of musical meaning; in other words a mechanical performance)
3	Child sings with some accuracy, starting in a different key than established, or modulates within the song	Child plays instrumental part with minor errors or gaps, recovers from mistakes
4	Child sings/chants melodic shape at significantly different pitch	Child plays instrumental part with some errors or gaps causing a clear interruption
5	Child sings/chants song with a different melodic contour than the song	Child strives to solve the tasks but plays mistakes such that the instrumental part is no longer recognizable
6	No meaningful evaluation possible (no or almost no singing)	No meaningful evaluation possible (no or almost no notes played)

Table 2. Rating scales used for annotating performance

Since we expected the input to be monophonic singing, this post-processing may seem superfluous. However, we realized that a certain amount of cross-talk was present in the recordings. This could lead to erroneous pitch detection and melody formation, since the pitch trajectories of the signal and the cross-talk are presumably quite close to each other. The main advantage of the high level tone objects over the frame-wise estimated pitch saliency values lies in the fact that tone objects assemble measures of past frames to establish spectral envelope information, as well as information about long term magnitude and pitch [5]. This means that with increasing duration of the sung note, pitch and magnitude estimates become more stable.

5.4 Audio Feature Extraction

Automatic melody transcription results in two representations of the main melody. First, the melody notes are characterized as discrete time events using MIDI parameters pitch, onset, and offset. Second, the course of the fundamental frequency is extracted using a time-resolution of 5.8 ms. Based on the transcription results, we computed a set of multi-dimensional audio features. The final feature vector has 138 dimensions. In the following sections, a selection of the used audio features is detailed.

Tonal features

Pitch Characteristics The first group of audio features characterizes the absolute pitch. The mean and the standard deviation were computed over the absolute pitch values to capture the average voice register and the pitch variability. Based on a histogram over all pitch class values[4], we computed the entropy as a feature. High entropy values indicate a flat distribution, which—assuming one diatonic scale for each melody tasks[5]—can indicate erroneous notes.

Melodic Contour The melodic contour describes the temporal curve of the absolute pitch values. The first features of this group are the pitch range and the number of transcribed notes. Then, the ratio of interval sequences with a constant interval direction as well a measure of the dominant direction (either ascending or descending) was computed. Based on [10], a simplified algorithm was developed, which segments a given melody into melodic arcs, which are consecutive pairs of ascending and descending note sequences. Features were computed by calculating the mean, standard deviation, minimum, and maximum over the arc lengths and pitch ranges.

Interval Characteristic Based on the interval values in semitones, measures for note sequences with constant pitch as well as chromatic note sequences were computed.

[4] The pitch class represents the absolute pitch but ignores the octave placement.
[5] No key changes were required in the melodies performed.

Melodic Intonation In order to characterize the melodic intonation, we computed a pitch class histogram of the frame-wise extracted fundamental frequency values as shown in Figure 4.

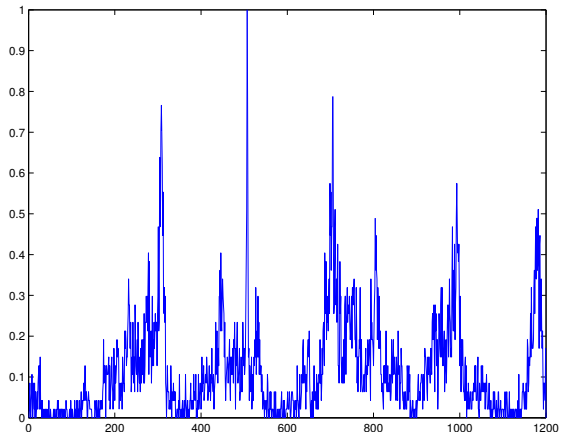


Fig. 4. Pitch class histogram n_{f_0} of a vocal recording

First, the f_0-values were converted from Hz to cent as

$$f_0^{[cent]}(k) = \left\{ 100 \cdot \left[12 \cdot \log_2 \left(\frac{f_0^{[Hz]}(k)}{440} \right) + 69 \right] \right\} \mod 1200 \tag{1}$$

The cent representation neglects the octave information. Then, a histogram $n_{f_0} \in \mathbb{R}^{1200}$ with a resolution of 1200 cents per octave was computed over f_0 and normalized to a maximum of 1.

As a reference, the equal temperament tuning was represented by a second histogram $n_{\mathrm{EQT}} \in \mathbb{R}^{1200}$ with

$$n_{\mathrm{EQT}}(k) = \begin{cases} 1 & k \bmod 100 \equiv 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

In order to compensate for global tuning deviations, the (reference) pitch class histogram n_{EQT} of the equal temperament tuning was circularly shifted until the cross-correlation between the two histograms n_{f_0} and n_{EQT} was maximized.

We derived one measure for precision of intonation as follows: For each of the highest 5 peaks in n_{f_0}, the distance to the closest peak in n_{EQT} was computed. These distances were weighted with the peak height in n_{f_0} and summed up. A precise f_0 intonation led to small deviations and hence to a small feature value. Additional features were computed from the size of the circular shift explained above and from the average peak width in n_{f_0}. Narrower peaks indicate a better intonation.

Melodic Fluctuation We computed features from the mean fluctuation (similar to a zero-crossing rate) and the mean intonation error for each note. In vocal performances, we often observed that the singer approaches the target note from a lower frequency at the beginning of the note. Similarly, at the end of the note, the fundamental frequency often dropped. Hence, we approximated the f_0 curve in the first and last 30 % of each note as a linear function and used the estimated function slopes as features.

Comparison to Reference Melody For each recording task, we encoded the target melody as a MIDI file and used it as the basis for comparison. Similarly to the melodic intonation features explained above, we computed a pitch class histogram n_{Ref} of the reference melody. The correlation coefficient between n_{f_0} and n_{Ref} was computed as feature.

Also, we compared both melodies on the basis of the number of notes and the edit distance (or Levenshtein distance).

Rhythmic features

Event Density We computed the mean and standard deviation of the number of notes per seconds as measures for event density and event density fluctuation.

Duration Characteristics The average note intonation, i.e., the ratio between note duration and inter-onset-interval, indicates whether notes were sung as legato or in a short and abrupt fashion. Furthermore, the occurrence of different note duration values were used as features.

6 Evaluation

6.1 Data set

In some of the recorded performances, the pupils did not sing or play at all. Since neither the automatic melody transcription nor the feature extraction could be applied to those recordings, they were removed from the data set before the annotation process started. The total numbers of annotated recordings were 617 vocal recordings and 664 recordings from the CMG app. Table 3 gives an overview over the data set for different rating classes.

6.2 Feature Selection & Classification

In order to reduce the dimensionality of the feature space, we applied an univariate filter feature selection method called Inertia Ratio Maximization using Feature Space Projection (IRMFSP) that was first proposed by Peeters and Rodet [13]. This feature selection algorithm is motivated by ideas similar to those of Fishers discriminant analysis. On each iteration of the algorithm we looked for the feature that maximizes the ratio of the between-class inertia to

Recording task	**Rating**				**Total**
	1	2	3	4	$\sum$
Instrument (CMG)	86	299	188	91	664
Voice	162	251	182	22	617

Table 3. Number of instrumental (CMG) and vocal recordings in the data set

the total-class inertia. To prevent chosen features from adding the same information in later iterations, all features were orthogonalized to previously selected ones. The algorithm could be stopped when the desired number of feature was chosen, or when the relative change of observed inertia ratio fulfilled predefined conditions. Using the feature selection, we reduced the feature space dimensionality to 50. We used a Support Vector Machine (SVM) with the Radial Basis Function (RBF) kernel as classifier. SVM is a binary discriminative classifier that attempts to find the optimal decision plane between the feature vectors of the different training classes [16].

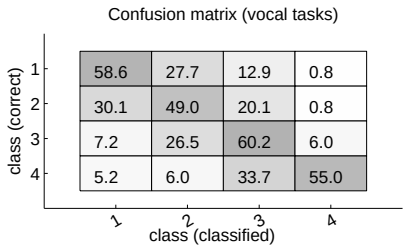
6.3 Evaluation procedure

We used a 20-fold stratified cross-validation and averaged the evaluation measures across all folds. We ensured that no recordings from the same pupil was used both as training and as test data within the cross-validation folds. Table 3 shows the number of unbalanced items among the 4 classes. Unbalanced class sizes may lead to a preference for larger classes in the trained SVM classification model which we avoided by employing sampling with replacement, in other words, we increased the number of class items for all classes to the number of class samples in the largest class by randomly sampling from the existing data. In the cross-validation, we made sure that the same items were never assigned simultaneously to the training data set and the test data set.

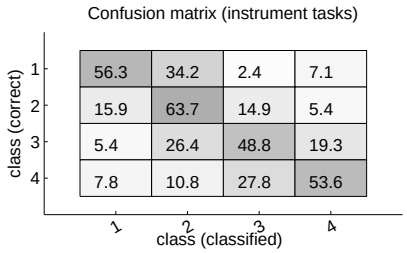
7 Results & Conclusions

7.1 Feature Selection

Table 4 shows the first 5 features that were selected by the IRMFSP feature selection algorithm for both data sets. For the vocal recordings, the selected features covered the total number of transcribed notes, the average pitch, two features that describe the note density, and the average note articulation. For the instrumental recordings, two of the selected features (Rank 2 and 4) measured the similarity between the transcribed melody and the target melody. The third feature captured the variability across all note durations. We do not have a plausible explanation for the feature on Rank 5 (probability of descending thirds).



(a) Vocal tasks (Mean class accuracy of 55.7 %)



(b) Instrumental tasks (Mean class accuracy of 55.6 %)

Fig. 5. Confusion matrix for the automatic performance assessment of vocal and instrumental CMG tasks.

7.2 Classification

As shown in Figure 5(b) and 5(a), similar mean class accuracy values of 55.7 % and 55.6 % were achieved for the vocal and instrumental (CMG) recordings. No comparative evaluation with other existing algorithm has been performed until now for two reasons. First, at this point in our research project, our data set cannot be published due to privacy restrictions. The data set published by Larrouy-Maestri et al. contains recordings of occasional singers. The expectable vocal quality is significantly higher compared to our vocal recordings and the classification task will not be comparable to ours. Second, none of the algorithms discussed in Section 4 were made public so we could not apply them to our data set. As expected, most confusions occurred between the adjacent rating classes. We assume that the overall classification performance was impeded by several factors:

- Incorrect transcription results. Even though we used an algorithm for predominant melody transcription, some of the male pupils sang in a comparably low register, so that some of the notes might have been transcribed incorrectly.
- It is possible that some features were melody-specific and immediately accessible to raters while they were not easy to identify by audio features

Rank	Vocal recordings	Instrument recordings
1	Number of notes	Flatness in onset-distribution histogram
2	Average fundamental frequency	Levensthein distance between transcribed and target melody
3	Average number of notes per second	Standard deviation of all note durations
4	Standard deviation of number of notes per seconds	Euclidean distance between pitch class histograms of transcribed and target melody
5	Average note articulation	Probability of descending third interval

Table 4. First 5 features, which were selected by the IRMFSP feature selection, for the vocal and instrumental (CMG) recordings

(e.g., phrasing at a certain place in the melody, particularly difficult interval). Think-aloud procedures with annotators will be undertaken to identify missing features.
- Sub optimal vocal recording conditions: especially when timid pupils sang very quietly, the singing of the neighboring pupils may have been transcribed erroneously instead. In order to minimize the cross-talk and distraction between neighboring pupils, we started to separate participants by wooden partition walls.
- In general, we observe significantly worse results than those reported by Molina et al. [12]. They assumed that the singing quality was good and the target melody was always complete. In contrast, we classified vocal performances of varying quality, ranging from quiet singing with spoken words to experienced vocal performances. The future testing of choirs may allow for larger data sets with lower variability in performance and improved possibilities for classifications.

8 Outlook

In order to add a rhythm factor to our current competency model with vocal and instrumental abilities, we started to record rhythm tasks using an adaptation of the presented CMG app. Large scale testing of up to 1000 pupils and 55 recordings per person is underway which will result in more annotated instances to be classified. Furthermore, we plan to improve classification results by adding an outlier detection scheme to the classification framework and by optimizing the applied audio features.

9 Acknowledgements

This research has been supported by the German Research Foundation (DFG BR 1333/10-1 and LE 2204/6-1).

References

1. Dittmar, C., Abeßer, J., Grollmisch, S., Lehmann, A., Hasselhorn, J.: Automatic Singing Assessment of Pupil Performances. Proceedings of the 12th International Conference on Music Perception and Cognition (ICMPC) and 8th Trinnial Conference of the European Society for the Cognitive Sciences of Music (ESCOM), Thessaloniki, Greece. (2012)
2. Dressler, K.: Sinusoidal Extraction Using an Efficient Implementation of a Multi-Resolution FFT. Proceedings of the International Conference on Digital Audio Effects (DAFx). Montreal, Quebec, Canada. (2006)
3. Dressler, K.: Pitch estimation by the pair-wise evaluation of spectral peaks. Proceedings of the Audio Engineering Society 42nd Conference on Semantic Audio (AES), Ilmenau, Germany. (2011)
4. Dressler, K.: An auditory streaming approach for melody extraction from polyphonic music. Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR). Miami, USA. (2011)
5. Dressler, K.: Towards Computational Auditory Scene Analysis: Melody Extraction from Polyphonic Music. Proceedings of the 9th International Symposium on Computer Music Modelling and Retrieval (CMMR). London, UK. (2012)
6. Hornbach, C. M., Taggart, C. C.: The Relationship between Developmental Tonal Aptitude and Singing Achievement among Kindergarten, First-, Second-, and Third-Grade Students. Journal of Research in Music Education, 53, pp. 322–331. (2005)
7. Jordan, A.-K., Knigge, J., Lehmann, A. C., Niessen, A., Lehmann-Wermser, A.: Development and Validation of a Competence Model in Music Instruction—Perception and Contextualization of music. Zeitschrift für Pädagogik. 58(4), pp. 500-521
   Lévêque, H., Schön, D., Giovanni, A., Morsomme, D.: The Evaluation of Singing Voice Accuracy: A Comparison between Subjective and Objective Methods. Journal of voice. (2012)
8. Larrouy-Maestri, P., Lévêque, H., Schön, D., Giovanni, A., Morsomme, D.: The Evaluation of Singing Voice Accuracy: A Comparison Between Subjective and Objective Methods. Journal of voice. (2012)
9. Larrouy-Maestri, P., Lévêque, H., Schön, D., Giovanni, A., Morsomme, D.: A comparison between subjective and objective methods for evaluating the vocal accuracy of a popular song. Proceedings of the 12th International Conference on Music Perception and Cognition (ICMPC) and 8th Trinnial Conference of the European Society for the Cognitive Sciences of Music (ESCOM), Thessaloniki, Greece. (2012)
10. Huron, D.: The Melodic Arch in Western Folksongs. Computing in Musicology, Vol. 10, pp. 3-23 (1996)
11. Mitchell, H. F., Kenny, D. T.: Open Throat: Acoustic and perceptual support for pedagogic practice. Journal of Singing, 64(1), pp. 429-441 (2008).
12. Molina, E., Barbancho, I., Gómez, E., Barbancho, A. M., Tardón, L. J.: Fundamental Frequency Alignment vs. Note-based Melodic Similarity for Singing Voice Assessment. Proceedings of the 8th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, Canada. (2013)

13. Peeters, G., Rodet, X.: Hierarchical Gaussian Tree with Inertia Ratio Maximization for the Classification of Large Musical Instruments Databases. Proceedings of the 6th International Conference on Digital Audio Effects (DAFx), London, UK (2003)
14. Persky, H. R., Sandene, B. A., Askew, J. M.: The NAEP 1997 Arts Report Card (Eighth-grade findings from the national assessment of educational progress). Washington: U.S. Department of Education. Office of Educational Research and Improvement, National Center for Education Statistics, NCES 1999-486r.
15. Salvator, K.: How can elementary teachers measure singing voice achievement? A critical review of assessments, 1994-2009. Update: Applications of Research in Music Education, 29(1), pp. 40-47, (2010)
16. Vapnik, V. N.: Statistical learning theory. Wiley New York (1998)