Automatic Singing Assessment of Pupil Performances

Christian Dittmar, Jakob Abeßer, Sascha Grollmisch,[*1] Andreas Lehmann, Johannes Hasselhorn[#2]

[*]*Semantic Music Technologies, Fraunhofer IDMT, Germany*
[#]*Hochschule für Musik, Würzburg, Germany*
[1]{dmr,abr,goh}@idmt.fraunhofer.de,
[2]{johannes.hasselhorn,ac.lehmann}@hfm-wuerzburg.de

ABSTRACT

Background

The rating of music competency in grade level school courses is our research interest. One aspect is the development of a systematic methodology for recording performances of pupils with regards to different singing tasks. Controlled and reproducible conditions for performance assessments are enforced through a proprietary software solution. The second aspect is the semi-automatic evaluation of the music proficiency, which will allow for large scale application in schools.

Aims

We aim to employ automatic analysis tools in order to model and predict the average of expert's summative ("overall") ratings as well as ratings of more specific vocal / musical attributes. This will lead to performance assessments of sung vocal performances.

The data set currently comprises 55 singing voice recordings of pupils at the age of 11 who were asked to sing the German national anthem. Three experts provided five-point-scale ratings of 19 attributes including the average expert rating as well as other attributes such as intonation, knowledge of melody, pitch range, and articulation. The averaged ratings served as ground truth for the performed experiment. The original melody is available as a reference in MIDI format. All recordings started with the first few notes of the melody played on a piano in order to provide a cue with respect to rhythm and tonality. However, the pupils had to finish the song without accompaniment.

Method

In order to model the expert ratings, we utilize methods from music information retrieval (MIR) and machine learning. We subject the singing voice recordings to an automatic melody transcription algorithm presented in Dressler (2011). This algorithm has two outputs that represent the sung melody: a discrete note sequence in MIDI notation as well as short-time frame-wise estimates of the fundamental frequency with a time resolution of 5.8 ms. Using a Query-by-Humming (QbH) algorithm similar to Ryynänen (2006) we tried to temporally align the performed melody and the reference melody.

Several performance assessment features are computed from the music transcription results in order to capture important aspects of the vocal performance. First, the sample correlation between different audio features and the ground truth ratings are investigated for all attributes. We could identify a first set of features, which show a significant correlation with several of the afore-mentioned attributes:

1) Pitch class histogram (PCH) difference
Based on the frame-wise fundamental frequency estimates, a pitch class histogram can be computed. The octave of the detected melody notes is discarded here; all pitch values are mapped to one octave. Here, we used a rather high frequency resolution of 1200 cent per octave (Kruspe, 2011). Similarly, we compute a pitch class histogram for the reference melody. Both pitch classes histograms are circularly shifted in such way that the cross correlation is maximized. Finally, the Euclidean distance between both aligned histograms is used as feature to indicate the tonal agreement between the performance and the melodic reference.

2) Peak-wise distance to an equal temperament PCH
First, we identify the five highest peaks in the PCH of the transcribed melody. The distance between these peaks and the closest PCH values that correspond to the equal temperament (multiples of 100 cent) are computed and normalized. This feature measures the precision of intonation.

3) Peak-width in the PCH
Based on the highest PCH peaks described in the previous feature, we compute the average peak width in cent as a feature. A well-intonated melody with stable fundamental frequency values results in well localized peaks exhibiting small width. In contrast, unstable intonation results in blurred and wider peaks.

4) Overall deviation from an equal temperament PCH
In order to measure the overall ratio of time frames with fundamental frequency values that do not coincide with the equal temperament, we filter the given melody PCH with a comb-filter with peaks at multiples of 100 cent.
Again, if the melody is well-intonated, the histogram will have a peak-like shape. Hence, the filtered histogram will contain most of the original values. The ratio between the sums over the filtered PCH vector and the original PCH vector is used as feature.

5) Entropy over a chroma histogram
Based on the note-wise pitch values returned from the sung melody transcription, we compute the pitch class of each note. All pitch class values are accumulated to a histogram. We compute the zero-order entropy as feature to characterize the sparsity of the histogram.
If only a subset of the 12 pitch classes as for instance the pitch classes associated to one diatonic key are sung, the entropy is lower than if out-of-key notes are frequently used. The latter

case would result in a more equally distributed pitch class histogram exhibiting a smaller entropy value.

6) Variability and drop of the fundamental frequency

By combining both the note estimates and the frame-wise fundamental frequency estimates, we can further examine how stable the intonation is. We compute two features based on the variance and the overall slope of the fundamental frequency over the course of each note. A similar approach was for followed for instance by Salomon et al. (2012).

Using this set of features, we performed several experiments using the regression algorithms multiple linear regression, stepwise regression, robust regression, partial least-squares regression, and support vector regression to automatically predict the given attributes. Due to the small amount of annotated recordings, we used a 20-fold cross-validation and averaged the prediction accuracy values over all folds.

Results

The correlation of the single features to the average expert rating score provides an indication of their quality with respect to performance assessment. Using the current feature set, we could achieve significant correlations ($p<0.05$) between the regressor's predictions and the expert annotations between $r=0.3$ and $r=0.6$ for approximately 80% of the attributes including the overall score. None of the compared regression methods showed superior performance, however, the multiple linear regression and the robust regression showed the best results.

Potential problems were identified during the experiments: Some pupils only sung parts of the national anthem and skipped to speaking the lyrics in between. In those cases, the melody transcription results were not reliable since the algorithm depends on a harmonic structure and can not identify noise-like spoken syllables and words.

Other pupils failed to follow the correct arrangement, which resulted e.g. in false melodic repetitions. These structural deviations impeded the alignment of the sung melody with the given reference MIDI file and resulted in erroneous feature values.

Conclusions

Our experiments show that the automatic modeling of expert ratings of children’s vocal performance is possible. However, it is obvious that more sophisticated features are needed. Future work will be dedicated to try a better alignment of the transcribed melody to the reference melody that is insensitive to tempo variations and can thus better model the sequence. In addition, rhythmic aspects will be investigated.

As a next step, a much larger data set will be assembled by recording pupils at several German schools. Furthermore, the recording conditions will be improved in order to ensure minimum interference signals.

Acknowledgements

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

Keywords

Hornbach Score, Music Competency Rating, Automatic Music Transcription

REFERENCES

Kruspe, A., Lukashevich, H., Abeßer, J., Großmann, H., & Dittmar, C. (2011). *Automatic Classification of Musical Pieces into Global Cultural Areas*. Proceedings of the 42nd Audio Engineering Society (AES) Conference: Semantic Audio, Ilmenau, Germany, 44-53.

Dressler, K. (2011). *Pitch Estimation by the Pair-Wise Evaluation of Spectral Peaks*. Proceedings of the 42nd Audio Engineering Society (AES) Conference: Semantic Audio, Ilmenau, Germany, 137-145.

Salomon, J., Rocha, B., & Gómez, E. (2012). *Music Genre Classification using Melody Features extracted from Polyphonic Music Signals*. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, Japan, 81-84.

Ryynänen, M. (2006). *Singing Transcription*. In Klapuri, A., & Davy, M. (Eds.), *Signal Processing Methods for Music Transcription* (1st edition, pp. 361-390). Springer Science + Business Media LLC