Automatic Competency Assessment of Rhythm Performances of Ninth-grade and Tenth-grade Pupils

Jakob Abeßer
Semantic Music Technologies
Fraunhofer IDMT, Germany
jakob.abesser@idmt.fraunhofer.de

Johannes Hasselhorn
Hochschule für Musik
Würzburg, Germany

Sascha Grollmisch
Semantic Music Technologies
Fraunhofer IDMT, Germany

Christian Dittmar
Semantic Music Technologies
Fraunhofer IDMT, Germany

Andreas Lehmann
Hochschule für Musik
Würzburg, Germany

ABSTRACT

In this paper, we introduce an approach for automated testing of music competency in rhythm production of ninth-grade and tenth-grade pupils. This work belongs in the larger context of modeling ratings of vocal and instrumental performances. Our approach relies on audio recordings from a specialized mobile application. Rhythmic features were extracted and used to train a machine-learning model which was targeted to approximate human ratings. Using two classes to assess the rhythmic performance, we obtained a mean class accuracy of 0.86.

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 and musical abilities is of interest. Music making is traditionally evaluated on an individual basis and results in testing procedures that can not be applied to large scale evaluations. One solution to this problem is simultaneous group testing.

The assessment of individual performances is an extremely time-consuming task. For example, a music teacher assessing five school classes, each consisting of 25 pupils performing for 5 minutes each, would have to listen to over 10 hours of recorded material. Therefore, a tool for both a simultaneous recording of all pupils as well as an automatic evaluation would be desirable when performing large-scale evaluation studies.

2. GOALS

Our goal was to measure the music making skills of pupils in German secondary school courses within the framework of competency modeling [1]. More precisely, we wanted to record vocal and instrumental music performances and develop a system for the automatic assessment of these recorded performances.

In order to add another facet to our current competency model that includes vocal and instrumental abilities [2], we started to record rhythm tasks using a special mobile application. Using an automatic rhythm analysis algorithm and annotations of the performance quality by music experts, we trained a statistical classification model of the experts' ratings.

3. PREVIOUS APPROACHES

In our own works [2, 3], we proposed how to estimate music competency of vocal and basic instrumental performance. Here, we devised a specialized mobile application that was used for (single voice) melodic input without requiring previous instrumental instruction [4]. To assess secondary pupils, we used the 5-point evaluation rubric originally developed by Hornbach and Taggart to assess elementary-age singers [5]. Its authors reported satisfactory inter-judge reliability values ($r = 0.76$ to $r = 0.97$).

Other authors outside education typically use tapping experiments and timing analysis rather than human raters [6]. Rhythmic synchronization and imitation has also been studied in [7, 8, 9].

4. NOVEL APPROACH

In our novel approach, we build upon the results and feedback obtained during our previous experiments dealing with vocal and instrumental performances. The focus on rhythmic competency made it necessary to develop new methodologies for testing and automatic evaluation, which will be

presented in this section.

4.1 Instrument Recording & Data Acquisition

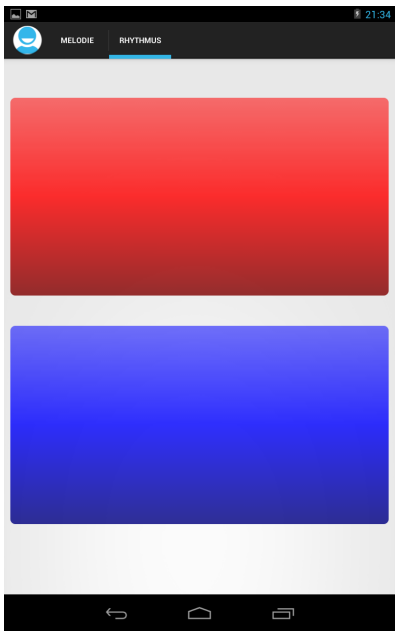

Figure 1. Screenshot of Colored-Music-Grid (CMG) app for rhythm tasks. The red and blue areas trigger the high and the low drum sound, respectively.

The participants in our experiments were 460 ninth-grade pupils. Each pupil worked at a separate workplace in groups of up to 25 pupils per classroom. Every workplace contained a laptop, a tablet computer, and a headset. The workplaces were separated using custom-made partition walls. The laptops were used to present the instructions of the different musical tasks consisting of text and scores to the pupil, and a headset was used to play audio examples and backing tracks. All laptops in the classroom were connected in a network, such that every task could be started simultaneously by the teacher on a separate computer.

A 7-inch tablet with a multitouch surface was used as musical instrument. Figure 1 shows a screenshot of the "Colored Music Grid" (CMG) app that we developed. Touching the red and blue areas triggers a high and a low percussive sound, respectively. Furthermore, this app provides a second mode that functions as a musical instrument for melody tasks as previously described in [4]. Each tablet was fastened on a holder above the laptop. The 27 rhythm tasks consisted of various one to two bar rhythm patterns, which were supposed to be performed alongside two different eight bar backing tracks.

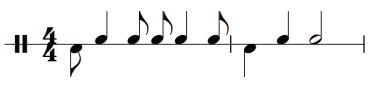

Figure 2. Example rhythm pattern.

Figure 2 illustrates an example rhythm pattern that was shown on the laptop screen. Additionally, the task instructions were given to the pupils via headset. In some cases, the instructions were reduced to a single sentence, in some instructions the backing track or the respective pattern was played as an audio example. This was followed by five seconds of silence, during which the pupils could practice the pattern. Next, the actual task instructions were played to the pupils starting with a one bar of count-in followed by the backing track for the current task. Each instrument performance was recorded as a separate audio track (44.1 kHz and 16 bit). The total dataset used in this paper consists of 8434 individual audio recordings.

4.2 Annotation

All recordings were evaluated by at least two out of 16 music students. For this evaluation, we used a six point ratings scale, which was adapted based on an established scale for the assessment of students singing performance [5]. For each recording, a rounded mean value was calculated from the two ratings. Depending on whether the task was rated by two or three raters, inter-rater consistency was estimated using Intraclass Correlations ICC(2,2) or ICC(2,3). Here, we used a two-way, random effect ICC, because two or three randomly selected assessors both rated all rhythm performances of one of the 27 tasks. ICCs varied between .67 and .93.

4.3 Audio Feature Extraction

Since we wanted to evaluate an audio recording of the rhythmic output of CMG, we needed to devise a suitable signal processing. In the following, we describe how the rhythm recordings were converted into an approximate transcription and what additional features were deduced from that.

4.3.1 Spectral Estimation

Based on a given audio recording of a rhythm performance, we first compute the Short Time Fourier Transform (STFT) using a blocksize of 2048 and a hopsize of 512. The given sampling rate (see Section 4.1) corresponds to a temporal resolution of approximately 10 ms. In the next section, we will explain how the drum envelope signals are extracted from the magnitude spectrogram $M(k, n)$ with k denoting the frequency bin and n denoting the time frame.

4.3.2 Drum Envelope Estimation

The CMG app uses two fixed samples for the low and the high percussion sounds without any dynamic changes. As can be observed in the upper subplot of Figure 3, the percussion spectra are well-separated in the magnitude spectrogram with barely any overlap in frequency ranges. Due to these idealized conditions, we apply a simple approach for drum envelope estimation. Based on the known frequency centroids of the drum sounds ($f_{\text{low}} = 1593\,\text{Hz}$ and $f_{\text{high}} = 2907\,\text{Hz}$), we extract the magitude envelopes $\hat{x}_{\text{low}}(n)$ and $\hat{x}_{\text{high}}(n)$ directly from the rows in $M(k, n)$ that correspond to the frequency centroids. The middle and lower subplot of Figure 3 illustrate two examples of magnitude envelopes of the two drum sounds for the excerpt shown in the spectrogram above.

In addition to the estimated drum envelopes, the correct reference drum tracks are given as MIDI files for each task. The MIDI files have two channels, one channel per percussion instrument. Based on these MIDI files, we generate two reference envelope functions $x_{\text{low}}(n)$ and $x_{\text{high}}(n)$ for each task by convolving the onset impulse function of each instrument with a Hanning window of 70 ms width.

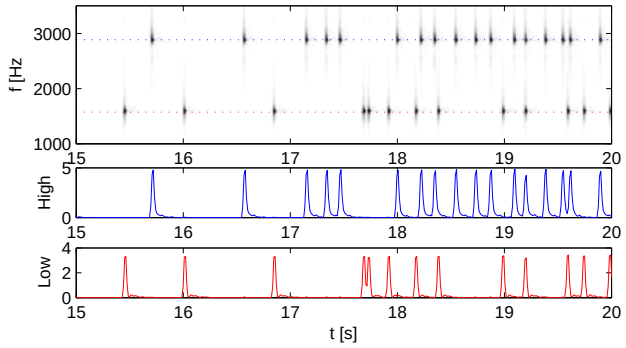

Figure 3. Excerpt from magnitude spectrogram $M(k, n)$ of drum recording (top). Time and frequency axis are given in seconds and Hz. Frequency centroids of high drum (blue) and low drum (red) are indicated as dotted horizontal lines. Resulting drum envelopes for high drum and low drum are given in the middle and lower subplot.

4.3.3 Envelope-based Features

All features described in Section 4.3.3 and 4.3.4 (denoted as F with a corresponding subscript) are extracted similarly for the low and the high percussive envelope, hence we omit the subscripts "low" and "high" for better readability.

The first group of features are extracted in order to compare the estimated drum envelope $\hat{x}(n)$ and the corresponding reference drum envelope $x(n)$ function. After $\hat{x}(n)$ and $x(n)$ are normalized to a maximum of 1, we compute the relative envelope energies

$$F_{\text{act}} = \frac{1}{N_{\hat{x}}} \sum_{n=1}^{N_{\hat{x}}} \hat{x}(n) \text{ and} \tag{1}$$

$$F_{\text{act,ref}} = \frac{1}{N_x} \sum_{n=1}^{N_x} x(n) \tag{2}$$

as features to measure the *drum activation*. $N_{\hat{x}}$ and N_x denote the number of items in $\hat{x}$ and x. Also, we use the *activation ratio*

$$F_{\text{actRatio}} = \frac{F_{\text{act}}}{F_{\text{act,ref}}} \tag{3}$$

as feature.

In the next step, we compute the cross-correlation $r_x(\tau)$ between $\hat{x}(n)$ and $x(n)$ to investigate to what extent both envelope functions coincide. Two features are obtained. The *envelope similarity* is measured by the maximum cross-correlation value $F_{\text{simEnv}} = \max_\tau r_x(\tau)$ and the *envelope synchronicity* is measured by the corresponding absolute shift value $F_{\text{syncEnv}} = |\tau_{\max}|$.

In addition to the cross-correlation, we count the number of local maxima in $\hat{x}(n)$ and $x(n)$ above a magnitude threshold of 0.05 as $N_{\max,\hat{x}}$ and $N_{\max,x}$. Here, the intuition is that local maxima in the envelope signal indicate individual note events. We compute features from the absolute difference over the local maxima number as

$$F_{\text{numPeakDiff}} = |N_{\max,x} - N_{\max,\hat{x}}| \tag{4}$$

and the ratio between the peak densities

$$F_{\text{peakDensRatio}} = \frac{N_{\max,x}/N_x}{N_{\max,\hat{x}}/N_{\hat{x}}}. \tag{5}$$

Finally, we compute a vector $d_{\max}$ with the temporal distances of adjacent local maxima in the envelope function $x(n)$. We compute the maximum, the mean, the variance, and the range over $d_{\max}$ as simple features to measure the amount of tempo fluctuation.

4.3.4 Features based on the Log-lag Autocorrelation

We compute the log-lag autocorrelation functions (LL-ACF) from $\hat{x}(n)$ and $x(n)$ as previously proposed in [10, 11, 12] over the tempo range of $10\,\text{bpm} \leq T \leq 600\,\text{bpm}$ with a resolution of 36 bins per octave. The LL-ACF represents a rhythmic pattern on a logarithmically-spaced lag axis and is comparable to the scale-transform [13]. The lags can be interpreted as reciprocals of the tempo. This means that small lags correspond to very high tempi, whereas lags to the end of the function correspond to extremely low tempi. The same rhythmic pattern played in different tempi result in similar LL-ACFs that are just shifted along the lag axis. The application of suitable distance measures for comparing LL-ACF has been discussed in [14, 15].

The LL-ACF of the estimated and reference drum envelope are denoted as $l_{\hat{x}}(T)$ and $l_x(T)$. Similarly as before, we compute the cross-correlation $r_l(\tau)$ between $l_{\hat{x}}(T)$ and $l_x(T)$ and take the maximum cross-correlation value $F_{\text{simLLA}} = \max_\tau r_l(\tau)$ and the corresponding shift $F_{\text{syncLLA}} = |\tau_{\max}|$ as features.

Next, we compute the energy sum

$$F_{\text{enSumLLA}} = \sum_T l_x(T) \tag{6}$$

and the ratio

$$F_{\text{enRatioLLA}} = \frac{\sum_T l_x(T)}{\sum_T l_{\hat{x}}(T)} \tag{7}$$

as features. As a next step, we extract the number of local maxima $N_{\max,l,\hat{x}}$ and $N_{\max,l,x}$ to describe both LL-ACF functions (we only consider maxima above 5 % of the highest peak). We compute the difference and the ratio between the number of peaks to measure the rhythmic similarity between the estimated and the reference drum envelope as

$$F_{\text{llacf,peakDiff}} = N_{\max,l,\hat{x}} - N_{\max,l,x} \text{ and} \tag{8}$$

$$F_{\text{llacf,peakRatio}} = \frac{N_{\max,l,\hat{x}}}{N_{\max,l,x}}. \tag{9}$$

Additionally, we compute average distance between adjacent local maxima in $l_x(T)$. In total, we obtained a 40-dimensional feature vector.

4.4 Automatic Modeling of Expert Ratings

We used a machine-learning approach to model the expert rating with the proposed audio features. The classifier model is trained based on expert ratings of a given training set. 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].

5. EVALUATION

5.1 Dataset

The dataset used in this paper consists of 8434 audio recordings with corresponding averaged performance ratings between 1 and 6. Figure 4 shows a histogram over the number of items for each class. Apart from class 1 and 6, the items are fairly well-balanced.

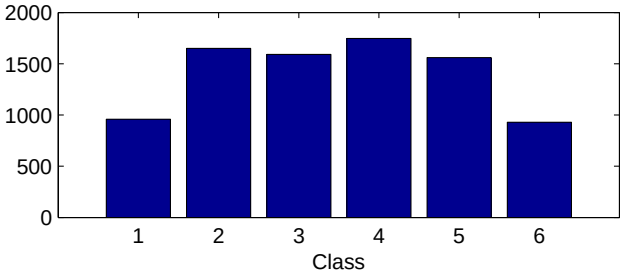

Figure 4. Number of items in the dataset for all six classes.

5.2 Experimental Procedure

5.2.1 Class Mapping

Based on the six-step rating scale discussed in Section 4.2, we investigated different class mappings to reduce the number of classes and thus to reduce complexity of the modeling task. In particular, we compared the 9 different class mappings as shown in Table 1. For each mapping, the six existing classes are mapped to two or three merged classes (denoted as C1, C2, and C3). Then, classifier models are trained based on the merged class annotations.

5.2.2 Cross-validation

For each mapping, we performed a 10 fold cross-validation and averaged the mean class accuracy over all folds. Since the class items are imbalanced as shown in Figure 4, we used a stratified cross-validation, i.e., we ensured that the proportion of items among different classes is kept approximately constant in each cross-validation fold. Furthermore, since we used Support Vector Machines classifier, we had to make sure that the number of items are balanced over all classes before training the model. Therefore, we used sampling with replacement, i.e., we increased the number of items in the smaller classes by randomly sampling from the existing data. At the same time, we ensured that similar items are never assigned as training and test data at the same time in the cross-validation procedure.

Table 1. Class mappings investigated in the evaluation experiment based on the original 6 classes. First column shows number of reduced classes. Second to fourth column show the original classes that are merged. Last column shows the mean class accuracy for the automatic classification (the highest value is emphasized in bold print). The last row gives the classification result if no class mapping is performed as reference.

#	# classes	Merged Classes C1	C2	C3	Mean Class Acc.
M1	2	1,2	4,5	-	**0.86**
M2	2	1,2,3	5,6	-	0.85
M3	2	1,2,3	4,5,6	-	0.80
M4	3	1,2	4,5	6	0.77
M5	3	1,2	3	5,6	0.69
M6	3	1,2	3,4	5	0.64
M7	3	1,2	3	4,5	0.63
M8	3	1,2	4	5,6	0.68
M9	3	1,2	3,4	5,6	0.68
M10	6	*No Mapping (6 classes)*			0.47

In each fold, we optimize the parameter values of γ and C of the RBF kernel function using a two-fold grid search as proposed in [17] with step sizes of 3 and 0.5 for the coarse and the fine grid search. Before each classifier training, the features are normalized to zero mean and a standard deviation of 1.

6. RESULTS & CONCLUSIONS

The last column of Table 1 summarizes the mean class accuracy values that we obtained for the different class mappings. It can be observed that by reducing the number of classes from six to two, the classification accuracy can be improved up to 0.86.

An initial experiment with the original six classes (without any class mapping) showed an accuracy of 0.47 and revealed strong confusions between adjacent classes, especially between and towards classes 3 and 4. Our results indicate that it seems beneficial to merge adjacent classes to more fuzzy categories such as good and bad. These categories can often be sufficient for an assessment of music performance.

Also, the removal of one of the medium classes (3,4) improved the classification results, as can be seen for instance when comparing M3 and M2 or M7 and M8, respectively.

Acknowledgments

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

7. REFERENCES

[1] A.-K. Jordan, J. Knigge, A. C. Lehmann, and A. Niessen, “Development and validation of a competence model in music instruction - perception and con-

textualization of music," *Zeitschrift für Pädagogik*, vol. 58(4), pp. 500–521, 2012.

[2] C. Dittmar, J. Abeßer, S. Grollmisch, J. Hasselhorn, and A. Lehmann, "Automatic singing assessment of pupil performances," in *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)*, 2012.

[3] A. Lehmann and J. Hasselhorn, "Assessing children's voices using hornbach and taggart's (2005) rubric," in *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)*, 2012, pp. 572–573.

[4] J. Abeßer, J. Hasselhorn, C. Dittmar, A. Lehmann, and S. Grollmisch, "Automatic quality assessment of vocal and instrumental performances of ninth-grade and tenth-grade pupils," in *Proceedings of the 10th International Symposium on Computer Music Modelling and Retrieval (CMMR)*, 2013.

[5] C. M. Hornbach and C. C. Taggart, "The relationship between developmental tonal aptitude and singing achievement among kindergarten, first-, second-, and third-grade students," *Journal of Research in Music Education*, 2005.

[6] J. Sowiński and S. Dalla Balla, "Poor synchronization to the beat may result from deficient auditory-motor mapping," *Neuropsychologia*, vol. 51, no. 10, 2013.

[7] W. C. Groves, "Rhythmic training and its relationship to the synchronization of motor-rhythmic repsonses," *Journal of Research in Music Education*, vol. 17, pp. 408–415, 1969.

[8] A. F. DeQuattro, "A comparison of the beat competency and rhythm pattern imitation ability of high school instrumental music students and high school dance students," Ph.D. dissertation, Boston University, Boston, 2013.

[9] C. Nombela, L. E. Hughes, A. M. Owenc, and J. A. Grahn, "Into the groove: Can rhythm influence parkinson's disease?" *Neuroscience & Biobehavioral Reviews*, vol. 37, pp. 2564–2570, 2013.

[10] C. Dittmar, M. Gruhne, and D. Gärtner, "Preprocessing methods for rhythmic mid-level features," in *Proceedings of the NAG/DAGA International Conference on Acoustics*, 2009, pp. 340–343.

[11] M. Gruhne and C. Dittmar, "Improving rhythmic pattern features based on logarithmic preprocessing," in *Proceedings of the 126th AES Convention*, Munich. Germany, 2009.

[12] M. Gruhne, C. Dittmar, and D. Gärtner, "Improving rhythmic similarity computation by beat histogram transformations," in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, Kobe. Japan, 2009, pp. 177–182.

[13] L. Cohen, "The scale representation," *IEEE Transaction on Signal Processing*, vol. 41, no. 3275-3292, 1992.

[14] T. Völkel, J. Abeßer, C. Dittmar, and H. Großmann, "Automatic genre classification of latin american music using characteristic rhythmic patterns," in *Proceedings of the 5th Audio Mostly Conference: A Conferencre on Interaction with Sound*, 2010.

[15] A. Holzapfel and Y. Stylianou, "A scale transform based method for rhythmic similarity of music," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 317–320.

[16] V. Vapnik, *Statistical learning theory*. New York: Wiley, 1998.

[17] C. Hsu, C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Taiwan, 2003.