Modeling memory for melodies

Daniel Müllensiefen[1] and Christian Hennig[2]

[1] Musikwissenschaftliches Institut,
Universität Hamburg, 20354 Hamburg, Germany
[2] Department of Statistical Science,
University College London, London WC1E 6BT, United Kingdom

Abstract. The aim of the presented study was to find structural descriptions of melodies that influence recognition memory for melodies. 24 melodies were played twice to 42 test persons. In the second turn, some of the melodies were changed, and the subjects were asked whether they think that the melody has been exactly the same as in the first turn or not. The variables used to predict the subject judgments comprise data about the subjects' musical experience, features of the original melody and its position in the music piece, and informations about the change between the first and the second turn. Classification and regression methods have been carried out and tested on a subsample. The prediction problem turned out to be difficult. The results seem to be influenced strongly by differences between the subjects and between the melodies that had not been recorded among the regressor variables.

1 Introduction

The main aim of the presented study was to find structural descriptions of melodies that influence recognition memory for melodies. A further aim was the exemplary comparison of statistical modeling approaches for data from psycho-musicological experiments.

Data have been obtained from a recognition experiment where melodies were presented twice to the experimental subjects. Some of the melodies were manipulated for the second presentation and subjects had to decide whether the melody had been changed or not. The experiment is described in detail in Section 2.

We tried to explain the judgments of the subjects with 19 predictor variables. This has been done by several classification and regression methods, which have been compared on a test set. The rating scale is ordinal, but we also carried out methods that predict variables on a nominal or interval scale. The prediction methods are described in Section 3 and some results are presented in Section 4. The best results are obtained by ordinal logistic regression and a random forest.

The prediction problem turned out to be hard. Even the best methods are not much superior to using the overall mean of the observations for prediction. In Section 5 we discuss some reasons. It seems that properties of the subjects and of the melodies that have not been captured by the explanatory variables play a crucial role.

2 The experiment

The primary motivation of the experimental design was to create a more realistic experimental scenario for a musical memory task than what is commonly used in similar studies (e.g. Eiting (1984), Taylor and Pembrook (1984), Dowling et al. (1995)). Thus, the design made use of musical material from a style that all subjects were familiar with (pop songs), it presented the objects to be remembered (melodies) in a musical context (arrangement), and the task required no specific musical training.

The sample consisted of 42 adults with a mean age of 29 and an average level of musical training that is similar to the German population. The musical material consisted of 36 MIDI polyphonic piano arrangements of existing but little known pop songs. The duration of each arrangement had been reduced to 50 seconds. From each song, a single line melody ("test melody", 15 seconds) had been extracted.

The task followed the "recognition paradigm" widely used in memory research (e.g., Dowling et al. (2002)). Subjects listened to the song arrangement and were played the test melody immediately afterwards. Then they were asked if the test melody has been manipulated or an exact copy of one of the melodies heard in the song. The ratings were done on a six-point scale encoding the subjects' decision and their judgmental confidence in three levels ("very sure no", "sure no", "no", "yes", "sure yes", "very sure yes"). The subjects were tested individually via headphones.

The idea behind the recognition paradigm is that correct memorization should result in the ability to detect possible differences between the melody in the song and the test melody. 24 melodies out of 36 (16 out of 24 for each subject) had been manipulated.

The following 19 predictor variables have been used:

- Time related factors:
  - position of the comparison melody in the song in seconds, in notes, in melodies, halves of song,
  - position of the manipulation in the test melody in seconds, in phrases of the melody, in notes of a phrase (or "no change"),
  - duration of the test melody in seconds, in notes.
- Musical dimensions of the melodies:
  - similarity of accent structures (as defined in Müllensiefen (2004)), overall similarity of the melodies (Müllensiefen and Frieler (2004)),
  - manipulation of the melody parameters rhythm, intervals, contour (or "no change"),
  - manipulation of the structural parameters range, harmonic function, occurrence of the repeated structure (or "no change").
- Musical background of the subjects: musical activity, musical consumption (summarizing scores have been defined from a questionnaire).

There are 995 valid observations. Subjects were asked whether they knew the song, and the corresponding observations have been excluded from the data analysis.

Particular features of these data are:

- The dependent variable is ordinal (though such scales have often been treated as interval scales in the literature). It is even more particular, because the six-point scale can be partitioned in the two halves that mean "I believe that the melody is manipulated" vs. "... not manipulated".
- The observations are subject-wise dependent.
- Some variables are only meaningful for the changed melodies. They have been set to 0 (all values for changed melodies are larger) for unchanged melodies, but this is doubtful at least for linear methods.

3 Prediction methods

Several prediction methods have been compared. The methods can be split up into regression methods (treating the scale as interval), classification methods (trying to predict one of six classes) and methods taking into account the nature of the scale. There were two possible codings of the six levels of the dependent variable, namely "1 $\simeq$ very sure changed",...,"6 $\simeq$ very sure unchanged" ("CHANGERAT") and "1 $\simeq$ correct prediction and very sure",..., "6 $\simeq$ wrong prediction and very sure" ("PQUALITY"), where the values 2, 3 indicate a correct answer by the subject but with less confidence in in his or her rating, and the values 4, 5 stand for a wrong answer with less confidence. For some methods, the coding makes a difference. One coding can be obtained from the other by using information present in the predictor variables, but it depends on the coding, which and how many predictor variables are needed. Not all methods worked best with the same coding. The following regression methods have been used:

- a linear model with stepwise variable selection (backward and forward, optimizing the AIC) including first-order interactions (products),
- a linear mixed model with a random effect for "subject" (variable selection as above),
- a regression tree,
- a regression random forest (Breiman (2001); default settings of the implementation in the statistical software R have been used for the tree and the forest).

The following classification methods have been used:

- a classification tree,
- a classification random forest,
- nearest neighbor.

Used methods that take into account the nature of the scale:

- ordinal logistic (proportional odds) regression (Harrell (2001), Chapter 13) with stepwise variable selection with modified AIC (Verweij and Van Houwelingen (1994)) and prediction by the predictive mean,
- a two-step classification tree and random forest, where first the two-class problem ("correct" vs. "wrong", PQUALITY coding) has been solved and then, conditionally, the three-class problem "very sure"/"sure"/"not sure".

The trivial methods to predict everything by the overall mean or, as an alternative, by the most frequent category, have been applied as well.

To assess the quality of the prediction methods, the data set has been divided into three parts of about the same size. The first part has been used for variable selection, the second part has been used for parameter estimation in a model with reduced dimension and the third part has been used to test and compare the methods. Methods with a built-in or without any variable selection have been trained on two thirds of the data. The three subsets have initially been independent, i.e., consisting of 14 subjects each. After obtaining the first results, we constructed a second partition into three data subsets, this time dividing the observations of every single subject into three about equally sized parts, because we were interested in the effect caused by the subject-wise dependence.

We used three performance measures on the test sample, namely the ratio of the squared prediction error and the error using the mean (R_1), the relative frequency of correct classification in the six-class problem (R_2) and the relative frequency of correct classification in the two-class problem (R_3, "change"/"no change", "correct"/"wrong", respectively). These measures are not adapted to ordinal data. A more problem-adapted loss function could be defined as follows: From a subject-matter viewpoint, it is "about acceptable" to predict a neighboring category. A prediction error of larger or equal than 3 can be treated as "absolutely wrong", and it is reasonable to assume a convex loss function up to 3. Therefore, the squared error with all larger errors set to 9 would be adequate. The results with this loss function should hardly deviate from R_1 without truncation, though, because most predictions have been in the middle of the scale, and prediction errors larger then 3 hardly occurred.

4 Results

Because of space limitations we only present selected results. We concentrate on R_1, which seems to be the most appropriate one of the measures described above. The results are given in Table 1. While the classification tree was better than the regression tree under R_2, both were dominated by the regression forest ($R_2 = 0.327$). Under R_3, the two-step forest (ignoring the second step)

Method	Partition 1 (independent)	Partition 2 (subject-wise)
Mean	1.000	1.000
Linear model	0.995	0.850
L. m./random effect	0.912	NA
L. m./r. e. (2/3 estimation)	0.890	NA
Regression tree	0.945	0.872
Regression forest	0.899	0.833
Reg. for. (subject ind.)	NA	0.762
Classification tree	1.062	NA
Classification forest	1.170	NA
Nearest neighbor	1.586	NA
Ordinal regression	0.912	0.815
Ord. reg. (all vars)	0.892	0.806
Two-step forest	1.393	NA
Two-step tree	1.092	NA

Table 1. R_1 results (all methods with optimal coding).

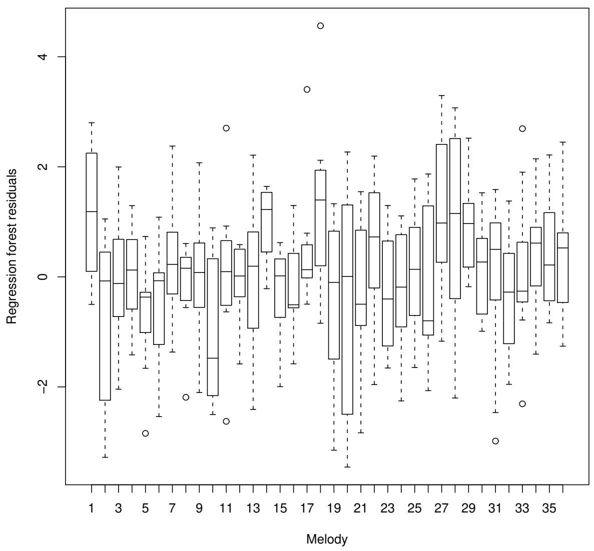

Fig. 1. Residuals (test sample) of regression random forest by melody.

was optimal ($R_3 = 0.670$), but not much better than the trivial guess "all judgments correct". Under R_2 and R_3, only a minority of the methods have been superior to the trivial "most frequent category" ($R_2 = 0.3$, $R_3 = 0.645$).

Under R_1 on the initial partition, the classification methods yielded values larger than 1 (i.e., worse than overall mean) and have been outperformed by the regression and ordinal methods. The regression forest (CHANGERAT coding) yielded a relatively good performance and provides useful information about the variable importance. The variable importance statistic "MSE increase if the variables would have been left out" for the random forest is more stable and therefore better interpretable than the selections of the stepwise methods because of the resampling character of the forest. The most important variables have been the overall melodic similarity, similarity of accent structures and the musical activity of the test persons. These variables are also among the four variables that appear in the regression tree.

Better results have been obtained by the ordinal regression on all variables without selection (while full models have been worse than models with reduced dimensionality for the linear models) and for a random effect linear model with variable selection,.

In general, the results are much worse than expected and demonstrate that the involved regression methods extract only slightly more information from the data than trivial predictors.

We suspected that this tendency is due to the fact that between-subjects differences dominate the judgments in a more complex manner than captured by the variables on musical background or the additive random effect of the mixed model. Therefore we repeated the comparison (without classification methods) on a partition of the data set where the same subjects have been present in all data subsets. The regression forest and the ordinal regression were the best methods in this setup (note that the overall mean, which is used as a reference in the definition of R_1, yielded a better MSE as well on this partition). By far the best result was obtained by a random forest including subject indicators as variables. The three variables mentioned above yielded again the highest importance statistics values.

The predictions have been improved on the second partition, but they still seem to be heavily dominated by random variations or influences not present in the predictor variables.

5 Further exploration and conclusion

We explored further the reasons for the generally weak performance of the methods compared to the trivial predictors. This led to two ideas:

- The familiarity of the structure of a melody (frequency and plausibility of melodic features) may play a key role. Figure 1 shows exemplary how the residuals of the random forest for the initial partition depend on the melody. A music-analytic look at the melodies with the highest positive residuals (1, 14, 18, 27, 28) reveals that they all include short and significant motifs of great "Prägnanz" (highly individual character), a feature that is hard to assess with quantitative methods.
- Different subjects show different rating behavior. It can be seen in Figure 2 that some subjects prefer less extreme ratings than others. The quality of the ratings varies strongly as well. These variations cannot be

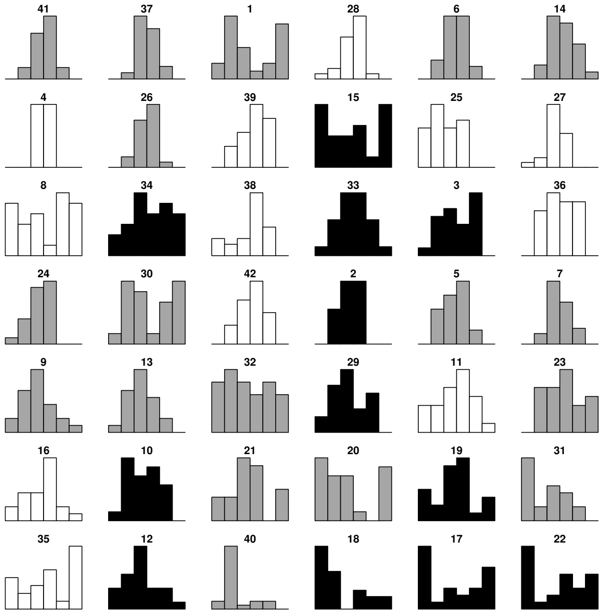

Fig. 2. Ratings (CHANGERAT coding) by subject. Every histogram gives frequencies for the ratings 1 to 6 over all melodies for one particular subject (numbers are subject indicators). Subjects are ordered according to their personal mean of PQUALITY (best raters on bottom right side, the worst raters - highest PQUALITY mean - are no. 41, 37, 1 and so on) and colored by musical activity (black $\simeq$ high activity, white $\simeq$ low activity).

fully explained by the musical activity and musical consumption scores or handled adequately by subject factors in the random forest or random effects. Figure 2 shows that high musical activity is related to a good rating quality, but the worst raters have medium values on the activity variable. Musical consumption (not shown) seems even less related to the subject differences. An idea to include these subject differences in the present study has been to perform a cluster analysis on the subject's rating behavior characterized by mean, variance and skewness of the two codings CHANGERAT and PQUALITY. A tentative visual cluster analysis revealed three clusters of particular subjects and a large "normal" group. We repeated the random forest on the second data partition including three cluster indicators. This yielded $R_1 = 0.766$. This result is biased because all observations were used for the clustering and the test sample has no longer been independent of the predictions. If done properly, the clustering should be performed on the first third of the data and the regression forest should be trained on the second third. But this would leave only 8 observations to cluster the subjects, which is not enough.

In general, the regression random forest seemed to be the most useful prediction method, especially for the assessment of the variable importance. The ordinal regression did a good job as well, but the main result of the study is the remaining large unexplained variation. This outcome suggests that the model is still lacking important predictors from the area of musical features. Such predictors should for example capture the “Prägnanz” of individual motifs.

It is interesting to see that in all applied models the two measures of melodic similarity and structure similarity are the variables with the largest explanatory potential. From a viewpoint of a cognitive memory model this means that the structural relation and the quantifiable differences between melody in the song and single line test melody is more decisive for memory performance than are experimental parameters (like the position of the target melody in the song or the duration of the different song parts) or information about the subjects’ musical background. In this sense, the results of this study shed some valuable light on the factors influencing recognition memory for melodies (even though the large amount of unexplained variance makes reliable indications of variable importance somewhat dubious).

Melodic features that may serve as further predictors are melodic contour, melodic and rhythmic complexity, coherence of melodic accents, and the familiarity of these features as measured by their relative frequency in a genre-specific database. The construction of new models making use of these novel melodic features are currently under investigation.

References

BREIMAN, L. (2001): Random forests. *Machine Learning, 45, 5–32.*

DOWLING, W. J., KWAK, S. and ANDREWS, M. W. (1995): The time course of recognition of novel melodies. *Perception & Psychophysics, 57(2), 136–149*

DOWLING, W. J., TILLMANN, B. and AYERS, D. F. (2002): Memory and the Experience of Hearing Music. *Music Perception, 19 (2), 249–276.*

EITING, M. H. (1984): Perceptual Similarities between Musical Motifs. *Music Perception, 2(1), 78–94*

HARRELL, F. E., jr. (2001): *Regression Modeling Strategies.* Springer, New York.

MÜLLENSIEFEN, D. (2004): *Variabilität und Konstanz von Melodien in der Erinnerung. Ein Beitrag zur musikpsychologischen Gedächtnisforschung.* PhD Thesis, University of Hamburg.

MÜLLENSIEFEN, D. and FRIELER, K. (2004): Cognitive Adequacy in the Measurement of Melodic Similarity: Algorithmic vs. Human Judgements. *Computing in Musicology, 13,147–176.*

TAYLOR, J. A. and PEMBROOK, R. G. (1984): Strategies in Memory for Short Melodies: An Extension of Otto Ortmann’s 1933 Study. *Psychomusicology, 3(1), 16–35.*

VERWEIJ, P. J. M. and VAN HOUWELINGEN, J. C.(1994) : Penalized likelihood in Cox regression. *Statistics in Medicine, 13, 2427.-2436.*