Perceptual evaluation of music similarity

Alberto Novello[1], **Martin F. McKinney**[1], **Armin Kohlrausch**[1,2]
[1] Philips Research Laboratories, Eindhoven, The Netherlands
[2] Human Technology Interaction, Technische Universiteit Eindhoven, Eindhoven, The Netherlands
{alberto.novello, martin.mckinney, armin.kohlrausch}@philips.com

Abstract

This paper presents an empirical method for assessing music similarity on a set of stimuli using triadic comparisons in a balanced incomplete block design. We first evaluated the consistency of subjects in their rankings and then the concordance across subjects. The concordance was also evaluated for different subject populations to assess the influence of experience of the subject with the musical material. We finally analysed subjects' ranking by the means of multidimensional scaling.

Similarity judgments were found to be rather concordant across subjects. Significant differences between musicians and non-musicians and between subjects being familiar or non-familiar with the music were found for a small number of cases.

Multidimensional scaling reveals a proximity of songs belonging to the same genre, congruent with the idea of genre being a perceptual dimension in subjects' similarity ranking.

Keywords: Perception, Music, Music Similarity.

1. Introduction

In the domain of Music Information Retrieval there has recently been an increasing interest in the automatic evaluation of music similarity. Although the task of identifying similar music often seems quite simple for a human listener, it is rather difficult to assess algorithmically or to represent it perceptually because of the multidimensionality involved in the human cognitive process.

The proposed theoretical models underline the multidimensional nature of perceived music similarity and stress the importance of the perceptual weighting of the various musical dimensions in this respect. Deliege's approach is an extension of the Gestalt theory applied to music [1]. The listener uses his/her prior experience to segment the musical piece and extracts features from every part. A weighted comparison between the features extracted from different parts can tell whether two parts are similar and in which respect. Okelford's zygonic theory is a more musicology oriented [2] alternative that tries to describe the feeling of musical derivation. This approach focuses on the relationships between notes (pitches, tempi, intervals, etc.) within a song, thus it seems more suitable for MIDI data than for raw audio signals. Cambouropulos' [3] unscramble algorithm/model also incorporates an important part of dimensional weighting but the nature of the weights is purely theoretical and its verification has been performed using just one song.

The multidimensionality of music similarity has mainly been explored through perception experiments [4]. Several studies [5, 6, 7] on this topic show the primary importance of the songs' tempo, genre and timbre in subjects' similarity ratings. In Chupchik's [5] experiments, the dominant dimensions used by subjects in similarity ratings were tempo, dominant instrument and articulation in a first experiment and tempo and genre in a second experiment. Dibben and Lamont [6] using a similar paradigm found that subjects primarily listened to "surface" features such as tempo, dynamics and articulation.

Considering the experimental framework in this domain, the research published so far appears very fragmented or too specific for general applicative interest based on a global representation of similarity [7-12]. As for the applications, few algorithms developed for assessment of music similarity between songs are linked to human perception.

From examining the literature in the domain of music similarity it appears that there is a lack of experimental verification of the theoretical perceptual models proposed [1, 2]. Examples of open experimental questions are: do subjects have a common perception of music similarity? What are the principal perceptual features used by subjects in rating similarity? Which features are most relevant? Is there a significant influence of musical experience (musicians/non-musicians, familiar/non-familiar musical material)?

2. Method

We use the method of triadic comparisons because it is a rather simple ranking procedure for subjects (compared to rating) and an efficient method to extract maximum information from a small set of stimuli. In the experimental session the subject is provided with three stimuli A, B and C and is asked to choose which pair is the most similar and which is the most dissimilar.

In the method of triadic comparisons, a *complete block design* is made with all possible comparisons (without symmetric repetitions). Having n stimuli, the number of triads is given by the formula:

$$b = \frac{n(n-1)(n-2)}{6} \quad (1)$$

A valid alternative is the *balanced incomplete block design* (BIBD) in which all possible pair-wise comparisons of stimuli occur λ times. The BIBD reduces the experimental time by a factor $\lambda/(n-2)$ compared to the complete design.

2.1. Experiment

Subjects were asked to listen to a set of three song-excerpts (triads) and choose for each comparison the most similar and most dissimilar of the three possible pairs. A questionnaire was also presented to each subject to evaluate subject familiarity with each stimulus, musical training, and general information (age, gender, etc.).

Our primary interest in the analysis of the whole complexity of popular music genres together with the limitations of experimental time lead to the choice of 18 audio excerpts stimuli spanning 9 genres: pop, rock, country, blues, jazz, heavy metal, hip hop, classical and funk.

One fast ($\geq$ 140 BPM) and one slow song ($\leq$ 100 BPM) were chosen for each genre. For each song, a representative section of 10 seconds was extracted from the chorus.

In the present experimental design, $\lambda = 2$ was chosen as a good compromise between experimental time and stimulus repetition. With 18 stimuli this factor gives a total of 102 triads in the BIBD. In addition, 10 triads were repeated in the design to examine subject consistency. The 10 repeated triads were chosen to span a range of difficulty as determined from a pilot experiment. The total number of triads was thus 112 and took about 50 minutes to complete. The obtained BIBD was randomly permutated to form 6 experimental designs to examine possible order effects. Thirty-six subjects participated: 18 musicians with at least 5 years of musical training and 18 non-musicians with no musical training (except compulsory low level school music courses).

2.2. Analysis

The data analysis can be separated in three parts: subject consistency, across-subject concordance and multidimensional scaling. In the first two parts Kendall's coefficient of concordance (KCC) is used to assess subject's concordance [13]. Kendall's W is defined as:

$$W = \frac{12S}{m^2(n^3-n)} \quad (2)$$

where S is the variance of the sum (across m judges) of ranks for n stimuli:

$$S = \sum_{j=1}^{n} (R_j - \bar{R})^2 \quad (3)$$

where R_j is the sum of judges' rankings for the j-th stimulus, and $\bar{R}$ is the average value for R_j (always equal to $\frac{1}{2}m(n+1)$).

2.2.1. *Within-Subject Analysis*

The 10 repeated triads were used to assess subjects' consistency. To do this, we calculated the KCC on the data from repeated trials for individual subjects. We used this measure to examine subject consistency in general and to compare subject consistency between musicians and non-musicians and between subjects familiar and unfamiliar with the music.

2.2.2. *Across-Subject Analysis*

The across-subject analysis was performed using only the 102 non-repeated triads in the BIBD. We first calculated KCC, across all the 36 subjects in the experiment for each of the 102 triads to evaluate general concordance of different subjects in their ranking and particular triad difficulty. Further analysis was conducted to find subjects whose pair ranking significantly reduced the overall concordance.

A third analysis divided subjects in two populations: musicians and non-musicians to assess across-subject concordance in each group. The bootstrap technique was applied to compare the two populations.

The final step of this analysis followed the same procedure. Subjects were divided depending on their familiarity with each triad into three groups: non-familiar, neutral and familiar. As not all triads had a sufficient number of subjects in both groups, familiarity and non-familiarity thresholds were adjusted to allow this verification on about half of the triads.

2.2.3. *Multidimensional Scaling*

After rejecting 8 outliers due to subject inconsistency (see below), we built a dissimilarity matrix of the remaining subjects', assigning the value 2 to the least similar pair, 1 to the middle pair and 0 for the most similar pair. We used the dissimilarity matrix as input to a multidimensional scaling (MDS) program calculation [14]. Given an input matrix of distances between points, the algorithm calculates the optimal positions of the points in an n-dimensional space.

In order to determine the optimal number of dimensions in the algorithmic calculation and plotting, we performed the MDS calculation with increasing number of dimensions (from 1 to 10) and recorded the stress value [14].

3. Results

3.1. Subject Consistency

For each subject, we calculated 10 concordance values W (one for each repeated triad), to measure subject consistency. The mean W values across 10 triads are plotted in Fig. 1.

We obtained the 5% significance level from the normal distribution of possible KCC values with 2 judges. The

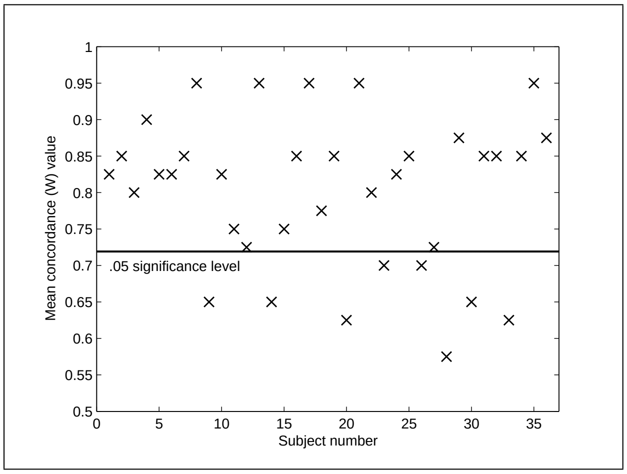

Figure 1. Mean concordance values for subjects on the 10 repeated triads to evaluate subject consistency

plot shows that on average 8 subjects (4 musicians and 4 non-musicians) are not significantly consistent within themselves on repeated triads'. These subjects were considered outliers and were rejected from the following part of the analysis. No significant differences in subject consistency were found between musicians/non-musicians or between subjects being familiar/non-familiar with the music.

3.2. Across-Subject Concordance

The $W_1, .., W_{102}$ values of KCC for each triad across all subjects are displayed in the Fig. 2. The data show general agreement across subjects: 97 triads from a total of 102 show rankings with significant concordance. Four of the five triads on which subjects show no concordance are formed by stimuli belonging to three different musical genres. The remaining triad is composed of two rock-excerpts and one heavy metal excerpt.

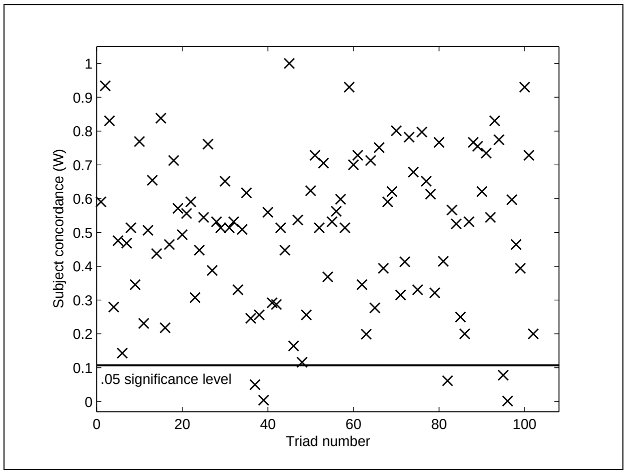

Figure 2. Across-subject concordance (W) per triad

Further analysis focused on the differences in distributions of the KCC values for musicians (W_m) and for non-musicians (W_{nm}). In this case, only 6 triads out of 102 show significant differences in concordance between musicians and non-musicians and on all these triads musicians show higher concordance than non-musicians. These triads (different from the triads shown in Fig. 2 on which subjects tend not to be concordant) contained stimuli belonging to three different genres. The difference between the distributions of the concordance for "non-familiar" subjects and "very-familiar" subjects has also been calculated for the triads that had a sufficient number of subjects in both groups. In this case again, only 5 triads out of the total non-rejected triads (about half of the original 102) show significant differences in concordance between familiar and non-familiar subjects and on all these triads the subjects non-familiar with the music perform more consistent than very-familiar subjects. All these triads again show three genres in their composition.

3.3. Multidimensional scaling

We used the MDS technique to represent subjects' rankings in a multidimensional metrical space. We computed the dependence of the stress factor (goodness of final plot to the original matrix) on the number of dimensions. It's common practice in literature [6] to consider a stress value between 0.15 and 0.1 as an acceptable limit value providing a reliable fit. In the present experiment, the optimum compromise was chosen to be 3 dimensions with a stress value of 0.157.

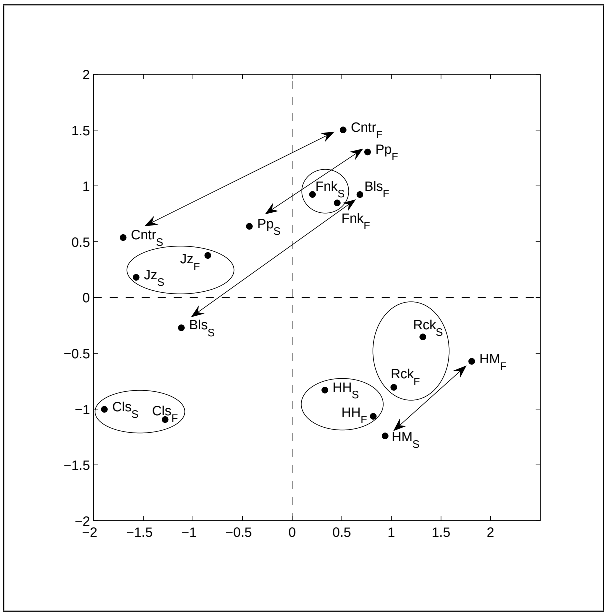

Figure 3. Excerpt positions in a 2-dimensional space, constructed with the ALSCAL algorithm to fit perceptual distances between excerpts (stress=0.244). Arrows and ellipses connect stimuli from same genre. Genre labels: Blues (Bls); Classical (Cls); Country (Cntr); Funk (Fnk); Hip Hop (HH); Heavy Metal (HM); Jazz (Jz); Pop (Pp); Rock (Rck). Subscripts: Slow Tempo (S); Fast Tempo (F).

Although the stress value in the case of 2D is too high for a good fit, we decided to show in Fig. 3 the results of the bidimensional MDS because of the difficulty of showing and interpreting a 3D plot in a paper format.

With only 2 dimensions it is already possible to see the effect of genre proximity for some genres (classical, funk, jazz, rock, hip-hop) while other genres (pop, blues, heavy metal, country) appear more dispersed.

In order to measure the existence of significant proximity in the stimuli depending on genre or tempo, we calculated the Euclidean distances of the three-dimensional positions determined as output by ALSCAL. The measure was conducted twice, comparing inter-tempo (songs classified in opposite tempo categories) with intra-tempo (songs in the same tempo category) distances, and inter-genre (songs belonging to different genres) with intra-genre distances (songs in the same genre). Significant differences were found only in the case of genre. Intra-genre distances were found to be significantly smaller than inter-genre distances.

4. Discussion

The experimental results show that despite 8 outliers, subjects show a common concordance on a large set of triads. Significant differences in consistency between musicians and non-musicians occurred only in a few triads, as was the case between subjects familiar and non-familiar with the music material. Nevertheless, the trend of the results on the triads that show significant differences needs some attention. In particular it seems interesting to understand which intrinsic characteristic of the 5 triads in Fig. 2 make the subjects' ranking less concordant. Nearly all these triads show three different genre stimuli in their composition. We think this characteristic might underlie an equal spacing of stimuli in a perceptual similarity space making it difficult for subjects to choose the most and least similar pair. The remaining triad appears consistent in this interpretative frame. It is composed of two rock excerpts and one heavy metal excerpt. If genre is one possible important dimension, in this case it also would impair the subject's ranking presenting two identical genre stimuli and one excerpt that belongs to a closely related genre.

We advance the hypothesis that musicians, who perform more concordantly on a fewer set of triads than non-musicians, have a more common approach to music interpretation than do non-trained listeners. For the case of music familiarity, a possible conclusion might be that subjects not familiar with the music (who show better concordance on few triads) make similarity judgments based more on the surface musical audio signal rather than on associated experience factors.

Through the use of a BIBD similarity ranking experiment and MDS of the resulting data, we have been able to represent subjects' "perceptual genre space": in 3-dimensions. Future work will extend this study to include a wider range of music.

5. Acknowledgments

We thank Jan Engel for the statistical support and Greg Dunn for his help running the experiment. This work is performed as part of a Marie Curie Early Stage Training grant (MEST-CT-2004-8201).

References

[1] Deliege I.,"Introduction, Similarity Perception - Categorization - Cue Abstraction", *Music Percept.*, 18(3), 233-43, (2001).

[2] Ockelford A., "On similarity, derivation and the cognition of musical structure", *Psychology of Music*, 32(1), 23-74, (2004).

[3] Cambouropoulos E. "Melodic Cue Abstraction, Similarity and Category Formation: A Formal Model", *Music Percept.*, 18(3), 347-370, (2001).

[4] McAdams, S. "Similarity, Invariance and Variation", *Annals New York Academy of Sciences*, (2001).

[5] Chupchik G. C., "Similarity and preference judgements of musical stimuli", *Scand. J. Psychol.* , 23, 273-282, (1982).

[6] Lamont A., Dibben, N., "Motivic Structure and the Perception of Similarity", *Music Percept.*, 18(3), 245-74, (2001).

[7] Eerola T., Jarvinen T., Louhivuori J., Toiviainen P., "Statistical Features and Perceived Similarity of Folk Melodies", *Music Perception*, 18(3), 275-96, (2001).

[8] Cahill M., "Melodic Similarity Algorithms - Using Similarity Ratings For Development And Early Evaluation", ISMIR 2005.

[9] Foote J., Cooper M., Nam U., "Audio Retrieval by Rhythmic Similarity", 265-266, ISMIR 2002.

[10] Aloupis G., "Algorithms for Computing Measures of Melodic Similarity", *Proc.* 15^{th} *Canadian Conference on Computational Geometry*, Halifax, 81-84, 2003.

[11] Allamanche E., Herre J., Hellmuth O., Kastner T., Ertel C., "A multiple Feature Model for Musical Similarity Retrieval", 265-266, ISMIR 2003.

[12] Berenzweig A., Logan B., Ellis D.P.W., Whitman B., "A Large-Scale Evaluation of Acoustic and Subjective Music Similarity Measures", ISMIR 2003.

[13] Kendall M., "Rank Correlation Methods", Charles Griffin, London (1975).

[14] Young F. W., Lewyckyj R., "ALSCAL User's Guide (5^{th} Ed.)", L.L. Thurstone Psychometric Laboratory, University of North Carolina, Chapel Hill, NC (1996).

[15] Kruskall J. B., "Multidimensional Scaling, Quantitative Applications in the Social Sciences" (Sage Publ. 1983)