

An ecological approach to multimodal subjective music similarity perception

Stephan Baumann

German Research Center for AI, Germany
www.dfki.uni-kl.de/~baumann

John Halloran

Interact Lab, Department of Informatics, University of Sussex, UK
www.cogs.susx.ac.uk/users/johnhall

Music information retrieval background. The perception and cognition of musically similar songs is a well-known research topic in the MIR community. There has been much work in the auditive area [e.g. Logan, 2001; Aucouturier, 2002]. More recent approaches investigate the influence of cultural factors on similarity through the use of metadata [Whitman 2002]. Possible bi-modal combinations of audio and cultural facets have been explored recently and [Baumann 2003] has added a third facet, lyrical similarity, to compute tri-modal subjective music similarity.

Experimental psychology and cognitive science background. We are interested in an 'ecological' approach to perception of musical similarity. This approach takes us out of the lab and into the actual worlds of music users. Being able to observe users' interactions with music free of experimental manipulation or direction could help us start to build accounts of perception of musical similarity based on cognition 'in the wild' [Hutchins, 1995]. These might have important implications both for further understanding of perception of music similarity, and designing the next generation of recommender systems.

Aims. We have two aims. The first is to establish a framework to conducting ecological studies. Second, to evaluate the relations between similarity facets, namely sound, lyrics, and cultural/style issues; and whether an ecological rather than a lab-based approach can lead us to new insights concerning music similarity perception in general, and these relations in particular.

Main Contribution. In order to allow the setting up of real-world experiments that allow us to study cognition 'in the wild' we have built a portable device, 'musicAssistant', which accesses a music database over WiFi in mobile usage scenarios. It allows users to directly interact with a recommendation engine. A virtual joystick controls the weights of 3 facets (sound, cultural and lyrics), each contributing to a combined music similarity measure. Our studies involve equipping music users with the musicAssistant and tracking the similarity decisions they spontaneously make in a variety of locations and social settings, at different times. These similarity decisions are then compared with those the same users make in a lab setting. Our research is beginning to reveal how perception of musical similarity might be impacted by social, temporal and geographic contexts in ways not predicted by literature based on lab-based studies. An ecological approach may give us new insights into the cognition of musical similarity and help us to understand the relation between the different facets involved.

Implications. The work of the MIR research community is shifting slowly towards user-oriented basic research and applied systems, as many recent contributions in core MIR conferences (ISMIR2003, DAFX2003) show. Cross-disciplinary work on different input data (raw audio, MIDI, metadata) and real-world user evaluation may provide means of addressing the open research areas in innovative ways.

Introduction

There is much current interest in creating man-machine interfaces that can support the digital distribution of music. The music information retrieval community is researching different ways of designing access to large archives in order to create new and effective means of delivering music to users according to their tastes. A critical issue is that of establishing computational models which approximate human perception of music similarity. These could help us to create systems that can automatically provide users with music they may like according to

their individual preferences. Different techniques have been used to inform such models, including content-based, webmining of cultural data, or combinations. Combined content-based and webmining approaches for timbral and cultural features have been developed by [Whitman and Smaragdis, 2002] and [Aucouturier and Pachet, 2002]. Other authors propose that models be informed by user studies, involving, for example, web-based collection of user ratings [Berenzweig et al, 2003], or lab-based studies [Allamanche et al, 2003]. Following on from this, our research asks how we might add to our understanding of perception of

music similarity through an 'ecological' approach. This means studying how people perceive music similarity in their normal lives beyond the artificial world of lab-based experiments. To this end we want to find new ways of observing users' interaction with our systems as they go about their everyday activities.

Cognition 'in the wild'

Cognition in the wild [Hutchins, 1995] means studying cognitive phenomena in the natural contexts in which they occur. For example, studies of marine navigation should take place on board ships. This approach relates to the insight that what people do in labs may not be 'ecologically valid' [Neisser, 1967]: experimental results may be artefacts of the lab situation, failing to represent people's behavior in the 'ecologies' of their normal lives.

While the lab-based approach can tell us about perception of music similarity, we feel it is also important to look beyond the lab and its artificial experimental setups, to music users' spontaneous perception of music similarity in real situations as part of their everyday lives. This ecological approach might reveal, for example, how perception changes with time, location, or activity, in ways which could have implications for how systems generate recommendations.

Multimodal music similarity

We have developed a trimodal global similarity measure which allows users to select a set of similar songs given an anchor song. This measure is realized as a weighted linear combination of three different local similarity metrics, namely sounds-alike similarity, similarity of lyrics and cultural or stylistic similarity:

$$S = wso * Sso + wly * Sly + wst * Sst$$

Sso : sounds-alike similarity

Sly : similarity of lyrics

Sst : similarity by style/cultural aspects.

In the research presented here, we wanted to see what sorts of decisions users would make in terms of these weightings in two different environments: (1) the lab; and (2) 'the wild': users' everyday lives as they pass from one

situation and activity to the next. In particular, we wanted to see if there are differences in terms of the salient cognitive dimensions on which music similarity is perceived. Might users use different weightings 'in the wild' than in the lab? And might these vary with different situations and activities?

Sound-alike similarity

Our users were presented with the means to choose similar songs on the basis of an MP3 archive. In order to compute a similarity matrix for this archive we follow a content-based approach. For each song timbral features are calculated which capture the overall sound by means of orchestration, prominent singers and typical studio or live recording characteristics. We compute mel frequency cepstral coefficients and a subsequent intra-song k-means clustering to generate a 'signature' for each song [Baumann, 2003].

In order to perform a similarity computation based on these signatures we use both the Earth Moving Distance, as suggested in [Logan2001] as well as a Gaussian Mixture Model as proposed in [Aucouturier, 2002]. Even working with a standard Nearest Neighbor (NN) classifier delivered interesting results for cross-genre recommendations of music 'sounding similar'.

Similarity of lyrics

For the experiments carried out here, we focused first on the subsymbolic level of lyrics. For each song in the music collection an according plain text file containing the lyrics is analyzed. The approach is known in information retrieval as a TFIDF weighted vector space model. The lyrics are represented as vectors. The components of the vectors indicate specific words; the value of a vector's component indicates the number of times the respective word occurs in the lyric represented. The similarity measure is the cosine-measure, which computes the angle between two vectors. In other words, those lyrics which share words will have high similarity. Typical results deliver similar songs from concept albums by the same artist (intra-album similarity, Example 1) or cross-artist similarity by topic (Example 2).

Example 1

P. Collins: ONE MORE NIGHT,
most similar:

P.Collins: You can't hurry love

P.Collins: Inside out

Example 2

C. Stevens: FATHER AND SON

most similar:

P.Collins: We're sons of our fathers

G.Michael: Father figure

Similarity by cultural/stylistic aspects

In order to compute cultural similarity we perform a Google search for reviews of the musical work of the artist under investigation. The first 50 pages are downloaded and pre-processed with a filtering stage to remove HTML formatting instructions and isolate the document text. A TFIDF-weighted vector space model is generated from the text, and additional part-of-speech tagging is performed to find semantically important words. As a result we achieve a 'cultural representation' or profile of the artist. It is a vector representation containing the most relevant distinguishing terms. On top of this vector representation the cosine measure is used to compute similarity between two entities. More details can be found in [Baumann, 2003] and [Whitman, 2002].

Music assistant

The multimodal music similarity measure was implemented and applied to a collection of 800 songs, compressed in MP3 format. To allow mobile access to this collection and recommendation functionality which uses the similarity measure we designed a web-based solution (Figure 1).

The website can be displayed on a small screen typical of PDA (personal digital assistant) or PIM (personal information management) devices. To allow the user subjective and interactive feedback we decided to include a virtual joystick which can be easily accessed using the pen of the PDA. The recommendation engine uses the similarity measure described. The position of the joystick has a direct influence on the

individual weights in the linear combination. In this way the individual users can select different settings and find their favorite combination, e.g.:

- center: $w_{so}=0.33, w_{ly}=0.33, w_{st}=0.33$
- sound: $w_{so}=1, w_{ly}=0, w_{st}=0$
- lyrics: $w_{so}=0, w_{ly}=1, w_{st}=0$
- style: $w_{so}=0, w_{ly}=0, w_{st}=1$



Figure 1. The music assistant consists of a wireless enabled handheld (PDA) and the functionality at server-side can be activated within a standard web browser (e.g. PocketExplorer for PocketPC) supporting flash plugin. The virtual joystick is implemented as a flash animation.

We were interested to see whether such combinations are influenced by research settings, either the lab or 'the wild'. If found, this would suggest that the cognitive dimensions on which music similarity is perceived might co-vary with setting, with implications for how portable recommender systems might work in real settings.

The technical design of the music assistant software and hardware supports ecological experiments. Devices with integrated wireless LAN features allow for mobile access to the webserver and the entire functionality. The logging mechanism at server-side allows for a non-intrusive observation of the users' actions. By using these features we designed an experiments to gain some insights about individual perception of music recommendations and differences between perception in the lab and 'in the wild'.

Experimental design

A group of 10 subjects roughly reflecting the current distribution of internet users in Germany was selected by age and education. We used a within-subjects design with two conditions: the lab, and 'the wild'. In each condition, each subject was asked to find the optimal joystick setting that would return an acceptable recommendation for a given anchor song. This position produces a particular trimodal weighting. Subjects were instructed that if they did not like the results of a given weighting, they could change that weighting immediately to produce an alternative result. They were also asked, if they liked the results of the weighting, to use the same weighting to select the next recommendation. If subjects ended up finding nothing, the system would generate another song without using any weighting.

In order to avoid learning effects different sets of songs were used for each condition and the order of presentation of conditions was randomized.

We gathered both quantitative and qualitative data. Quantitative data was generated by an automatic server-side logging mechanism which collected the individual joystick settings that led to the user-intended results. Qualitative data consisted of observations and unstructured interviews.

Experimental results

We carried out quantitative and qualitative analysis of our results.

Our findings from the quantitative analysis show that regardless of condition there is a general preference for a particular trimodal music similarity measure: the typical setting reflects an average importance of sound=0,41, style=0,36, lyrics=0,22.

As well as these general findings, we found that there were some differences across the two conditions. The wild environment led to a slight increase of the style facet which may be due to the fact that the lab environment was silent and allowed the users to concentrate on the sound facet more easily than 'in the wild'.

ALL	Sound	Style	Lyrics
Median	0,4	0,35	0,12

Avg	0,41	0,36	0,22
StdDev	0,27	0,27	0,27
LAB	Sound	Style	Lyrics
Median	0,48	0,35	0,10
Avg	0,44	0,34	0,21
StdDev	0,26	0,25	0,27
WILD	Sound	Style	Lyrics
Median	0,37	0,35	0,16
Avg	0,38	0,38	0,24
StdDev	0,27	0,29	0,26

Table 1. Experimental results.

Furthermore, there were individual differences within our group of ten subjects which were, again, consistent across the two conditions. We identified different types of user according to a qualitative analysis of the recorded statements after the experiments and a cross-check against the quantitative analysis.

Type 1 users' main interest was in the cultural/stylistic facet. They explicitly stated that they ignored the lyric facet. Typical quotes included:

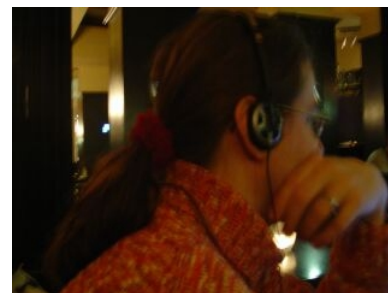


Figure 2. Type1 subject 'in the wild'.

- "I am not interested in lyrics in general, because my English is too bad to understand something" (remark: German native-speaker)
- "I know all these recommendations from title, I don't have to listen to it, which saves my time"

Type1	Sound	Style	Lyrics
Median	0,21	0,78	0,01
Avg	0,28	0,67	0,04
StdDev	0,24	0,20	0,06

Table 2. Results from a Type 1 user.

Type 2 users were mainly interested in the sound of a song. This was at least partly due to the capabilities of the device:



Figure 3. Type 2 subject 'in the wild'.

- "The sound similarity is really impressive, on its own it is able to deliver songs I did not know before which fit my taste"
- "After a while I got used to this and played around just like browsing into music"
-

Type2	Sound	Style	Lyrics
Median	0,54	0,15	0,00
Avg	0,60	0,20	0,19
StdDev	0,28	0,21	0,27

Table 3. Results from a Type 2 user.

Type 3 users found lyrics to be interesting. This facet acted as a bias that helped them to discover previously unknown songs in novel ways.

- "I used the lyric facet to get more interesting results beyond the mainstream opinion"
- "by adding a lyrics a bit, the results get more interesting"
- "some German hiphop recommendations sound totally different but the lyrics are funny"

Type3	Sound	Style	Lyrics
Median	0,30	0,34	0,34
Avg	0,29	0,31	0,38
StdDev	0,24	0,28	0,33

Table 3. Results from a Type 3 user.



Figure 4. Type3 subject 'in the wild'.

Conclusion

In this paper we undertook research to find out whether an ecological approach to perception of music similarity might reveal how it differs in 'the wild' to lab-based settings. This might have important implications for the design of mobile music recommendation systems.

Our results show the opposite: whether in the lab or 'in the wild', all users (including the types we identified) were consistent in their preferred configuration of the trimodal similarity measure. This proved robust across situations, locations and times. What this suggests, on the face of it, is that there is no great difference between the two settings, and so no pressing need to implement an ecological approach to multimodal subjective similarity perception. However, an issue arises about how far we were able to create truly 'wild' settings. The non-lab settings we were able to create were constrained by technical problems. Critical among these were (a) the absence of stable WLAN hot spots, and (b) the difficulty of accessing legally available, large-scale musical content of interest. The first of these issues, in particular, meant that it was difficult to track behavior across times and activities, since mobility was highly restricted. However, even given these problems, we found evidence that ambient noise has an effect on how weightings are set. Before the ecological approach is dismissed, different findings, including further information about different kinds of ambient noise, could result from being able to access robust WLANs as well as large music corpuses. Both of these technical problems remain challenging. Simulations may be needed in order to find out about how the coming generation of mobile music recommendation systems work in people's real lives.

Acknowledgments. We wish to thank the *Stiftung Rheinland-Pfalz für Innovation* for supporting this work.

References

- Aucouturier, J. and F. Pachet. , "Music Similarity Measures: What's the Use?", in Proceedings of the ISMIR 2002, Paris, France, pp. 157-163, 13-17 October, 2002.
- Allamanche et al. , "A multiple feature model for musical similarity retrieval", in Proceedings of the ISMIR 2003, Baltimore, USA, October, 2003.
- Baumann, S., "Music Similarity Analysis in a P2P Environment", in Proceedings of the 4th European Workshop on Image Analysis for Multimedia Interactive Services, London, UK, April 2003.
- Baumann, S. and O. Hummel. *Using Cultural Metadata for Artist Recommendation*. In Proceedings of the Third International Conference on Web Delivering of Music, Leeds, UK, September 2003.
- Berenzweig, A. et al. *A large-scale evaluation of acoustic and subjective music similarity measures*. In Proceedings of the ISMIR 2003, Baltimore, USA, October, 2003.
- Hutchins, E. (1995) *Cognition in the Wild*. MIT Press.
- Logan, B., "A Content-Based Music Similarity Function", Technical report, Compaq Cambridge Research Laboratory, June, 2001.
- Neisser, U. (1976) *Cognition and Reality*. Freeman.
- Whitman, B., Smaragdis, P., "Combining Musical and Cultural Features for Intelligent Style Detection", in Proceedings of the ISMIR 2002, Paris, France, 13-17 October, 2002.