Towards the development of preference models accounting for the impact of music production techniques

Janto Skowronek[1], Lukas Nagel[2], Christoph Hold[2], Hagen Wierstorf[3], Alexander Raake[1]
[1] *Technische Universtität Ilmenau, Institut für Medientechnik, Email: janto.skowronek@tu-ilmenau.de*
[2] *Technische Universtität Berlin*
[3] *Filmuniverstität Babelsberg KONRAD WOLF*

Introduction

The conventional approach to assess the quality of a system from the end user's perspective is based on the paradigm that the system under test is degraded compared to an ideal system. Depending on the test protocol, this ideal system is explicitely presented as a reference condition (e.g. [1], [2], [3] Annex D) or it is assumed to be present in the user's mind in form of an internal reference (e.g. [3] Annex A & B, [4]). For spatial audio systems in the musical context that go beyond stereo sound reproduction, however, it has not yet been possible to establish such a concept of an optimum reference.

For that reason, it appears to be less appropriate to investigate the user's appreciation of spatial audio systems by means of quality ratings using one of the well established scales and test protocols [1, 2, 3, 4]. Instead, pairwise preference ratings appear to be a promising alternative, as they simply ask for a number of stimuli pairs which of the two stimuli is preferred without requiring the participant to consider an optimum spatial audio system.

While the paired comparison method is a good candidate to collect preference data for spatial audio systems, the effort of conducting such test series is quite large. For that reason, a model is desired that is able to predict such human preference data. The present paper discusses a proof-of-concept study on developing such a model, using the data obtained in a related study [5].

However, the related study investigated in particular the impact of the music production techniques, i.e. the mixing proccess, on the preference ratings. Since the perceptual data used for this paper is taken from [5], the presented proof-of-concept model actually accounts more for the impact of the mixing parameters than for the impact of the spatial sound reproduction system.

Perceptual data

For the listening experiment, Hold [5] generated the stimuli pairs by modifying four different mixing parameters: compression, equalization, reverb, positioning. In addition, Hold generated also a fifth category – vocal processing – for which he modified compression, equalization and reverb only on the vocal tracks. Hold chose three or four different instantiations for each mixing parameter, and included one reference mix for wavefield synthesis reproduction and one for stereo reproduction. Thus, the total set of stimuli comprised 19 conditions, see Table 1.

To limit the overall experimental effort, 103 of the 171 possible stimuli pairs ($\approx 60\%$) were presented during the listening experiment.

Table 1: The 19 stimuli used in the perceptual experiment.

Name	Mixing parameter	Short description
WFS	—	WFS reference mix
STE	—	Stereo mix
C_M	Compression	Half of the gain reduction of the compressor, compared to the WFS reference mix
C_MM		Compressor switched off
C_P		Double of the gain reduction of the compressor, compared to the WFS reference mix
E_M	Equalization	Half of the amount of cutting or boosting in the EQ filters, compared to WFS mix
E_MM		EQ filters switched off
E_P		Double of the amount of cutting or boosting in the EQ filters, compared to WFS mix
R_M	Reverb	Half of all reverb return signals, compared to WFS mix
R_MM		All reverb return signals switched off
R_P		Double of all reverb return signals, compared to WFS mix
P_M	Positioning	Shifting of foreground elements towards the center: wider than stereo, narrower than WFS
P_MM		Shifting of foreground elements towards the center: within the two stereo speaker positions
P_P		Slight spreading of foreground elements farther apart and away from the center
P_PP		Extreme spreading of foreground elements farther apart and away from the center
V_M	Vocals	Reduced processing of vocals: half of Compression, EQ, and Reverb settings compared to WFS mix
V_MM		Processing of vocals switched off
V_P		Increased processing of vocals: double of Compression, EQ, and Reverb settings compared to WFS mix
V_Pb		Increased processing of vocals, variant: half of Compression and EQ settings compared to WFS mix

Ground-truth data aquisition

Figure 1 visualizes the processing stages to obtain the ground-truth data for the later modeling part. Starting point is a stimulus pair A and B, which is presented in the listening-only test (see [5] for details) to 41 test participants, who are asked to indicate which of the two stimuli they prefer.

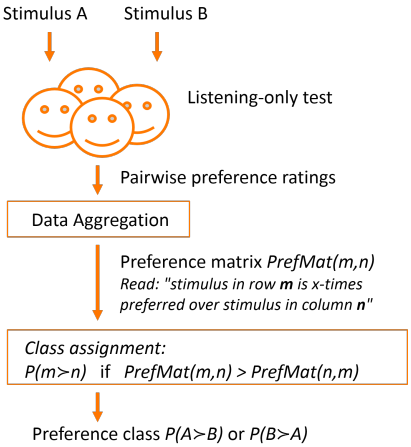

Figure 1: Data aquisition processing stages.

These pairwise preference ratings are then aggregated over all test participants and stored into a preference matrix *PrefMat(m,n)*. The elements in that matrix contain the numbers how often a stimulus was preferred over the other, and it should be emphasized that this matrix does not need to be symmetric. Table 2 shows the preference matrix computed for the data used.

Finally, the ground-truth data is obtained by assigning one of the two preference classes $P(A \succ B)$ or $P(B \succ A)$ to each stimulus pair (A, B), depending on which of the two stimuli is more often preferred. This can be computed from the preference matrix by:
$P(m \succ n)$ *if* $PrefMat(m,n) > PrefMat(n,m)$.

Model structure

Since the ground-truth data is considered as classes – class 1: $P(A \succ B)$, class 2: $P(B \succ A)$ – that are assigned to each stimulus pair, the modeling approach (see Figure 2) is based on a conventional classification algorithm, consisting of a feature extraction and pattern recognition stage. These two stages are extended with a difference computation stage, as the pattern recognition stage has the task to decide for a class based on *comparing* two input signals. Those three stages are described in more detail now.

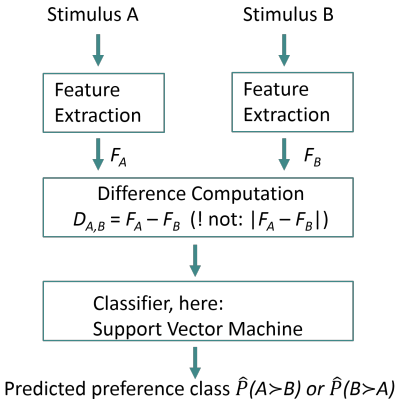

Figure 2: Model structure.

Stage 1: Feature Extraction

Since the stimuli differed in four mixing parameters (compression, equalization, reverb, positioning), we hypothesized that a powerful feature set should comprise four different feature types, each dedicated to characterize one mixing parameter. Furthermore, to minimize overtraining effects given the rather small data set of 206 data points (103 stimuli pairs, 2 difference feature vectors per stimuli pair, see below), we aimed at a number of three to four features per feature type. The following text provides a short description of the four feature types comprising in total 15 features.

LDR (3 features, characterizing *Compression*):
Skovenborg [6] proposed a feature LDR that describes microdynamic behavior in music by computing the 95 percentile of difference values (in dB) between a slow loudness function (3s integration time) and a fast loudness function (25ms). Since this feature showed a promising correlation with the perceived dynamics [6], we chose this feature as a candidate for characterizing compression, whereas we adopted Skovenborg's computation by calculating the slow and fast loudness signals from the gammatone filterbank outputs of the Two!Ears Auditory Frontend [7]. In addition, we also computed the Pearson correlation coefficient and the Root Mean Square Error between fast and slow loudness signals.

SPEC (4 features, characterizing *Equalization*):
Since equalization means to modify the relative amount of signal energy in different frequency areas, spectral features are apparently good candidates. In his master's thesis, Nagel [8] investigated the potential usefulness of six spectral features, which were computed using the Two!Ears Auditory Frontend, for a first subset of the perceptual data (21 test participants instead of 41). From those six features we took the four most promising features: Decrease, Variation, Entropy, and Irregularity.

VDS (4 features, characterizing *Reverb*):
Van Dorp Schuitman et al. [9] developed a model to estimate four room acoustic parameters from binaural input signals: reverberance, clarity, apparent source width, and listener envelopment. Hypothesizing that applying reverb in a music mix leads to a similar perception than the perception of the acoustical properties of a real room, we considered the four parameters as good candidates for the characterization of the reverb mixing parameter. Note that the features were computed with the original implementation of the van Dorp Schuitman model, since the model was not yet integrated into the Two!Ears framework at the time when we conducted this study.

LOC (4 features, characterizing *Positioning*):
Hearing different music elements in a mix at different positions means that they have different localization cues. Therefore, we decided to exploit the ITD and ILD cues extracted from the Two!Ears Auditory Frontend. Motivated by the hypothesis that the different positioning mixes lead to different variations of ITD and ILD cues over time and frequency, we computed for ITD and ILD two of the four combinations of the mean and standard

Table 2: Preference matrix for the 19 different stimuli. Read the matrix as follows: Stimulus in row m is x-times preferred over stimulus in column n. The main diagonal is not defined (indicated by –). Empty cells mean that the corresponding stimulus pairs have not been tested in the listening experiment.

	WFS	STE	C_M	C_MM	C_P	E_M	E_MM	E_P	R_M	R_MM	R_P	P_M	P_MM	P_P	P_PP	V_M	V_MM	V_P	V_Pb
WFS	–	153	20	30	27	19	20	21	20	24	24	25	25	24	23	20	20	25	28
STE	52	–	8	18	11	12	14	11	6	10	17	6	11	16	16	14	13	18	15
C_M	21	33	–	23	21														
C_MM	11	23	18	–	18		21	16		14	20		23		24		15	24	22
C_P	14	30	20	23	–		25	23		27	21		24		21		25	27	22
E_M	22	29				–	27	23											
E_MM	21	27		20	16	14	–	17		18	25		23		21		20	25	23
E_P	20	30		25	18	18	24	–		25	20		22		19		22	23	22
R_M	21	35							–	22	20								
R_MM	17	31		27	14		23	16	19	–	24		22		20		20	28	22
R_P	17	24		21	20		16	21	21	17	–		22		22		23	25	22
P_M	16	35										–	22	27	20				
P_MM	16	30		18	17		18	19		19	19	19	–	23	22		20	22	22
P_P	17	25										14	18	–	24				
P_PP	18	25		17	20		20	22		21	19	21	19	17	–		17	19	17
V_M	21	27														–	23	30	
V_MM	21	28		26	16		21	19		21	18		21		24	18	–	23	23
V_P	18	25		19	16		18	20		15	18		21		24	11	20	–	17
V_Pb	11	24		17	17		16	17		17	17		17		22		16	24	–

deviation (STD) across time and frequency (i.e. Auditory Frontend bands): Mean over bands and STD over time, STD over bands and STD over time.

Stage 2: Difference Computation

In order to enable the algorithm to identify which of the two stimuli would be more preferred, a measure is required that keeps the sign. For that reason, the measure computed here is not a distance measure but the simple subtraction of the feature vectors belonging to the two stimuli $D_{A,B} = F_A - F_B$.

This requirement of keeping the sign, however, implies that the order of stimuli put into the model would influence the result, as $D_{A,B} = F_A - F_B \neq D_{B,A} = F_B - F_A$. Since the model does not know, which stimulus is sent to which model input, the classification algorithm needs to be able to predict the preferred stimulus for both cases $D_{A,B}$ and $D_{B,A}$. That means, the pattern recognition stage needs to be trained with both cases, i.e. the difference feature vector $D_{A,B}$ with the corresponding ground-truth preference class $P(A \succ B)$ and the inverted difference feature vector $D_{B,A}$ with the corresponding inverted ground-truth value $P(B \succ A)$.

Stage 3: Pattern Recognizer

Support Vector Machines (SVM) are used as pattern regocnition method, since the model task as a two-class problem, and SVMs have been successfully used for audio and music classification problems [10].

Model training and evaluation

We decided to test a number of models using different combinations of the four feature types in order to investigate their contribution to the model performance. The idea is to compare models in which either all four feature types are used ($M1$), in which only one feature type is used ($M2$ to $M5$), and in which all feature types except one are used ($M6$ to $M9$).

As evaluation method, we opted for the bootstrap method [11] – or more precisely a variant of it used in other works [12, 13] – in which the model is trained and tested multiple times. In each repetition, the data is randomly split into a training set, on which the model is trained, and a test set, on which the model is evaluated. Using 100 such repetitions and choosing for a training to test data ratio of 80/20, we computed as performance measures the mean and 95% confidence intervals of the classification performance.

Results and discussion

Figure 3 shows the model performance for the nine tested models. A first result is that the model performance is in all cases above the chance level of 50%. Thus, this study suggests that it is in principle possible to predict pairwise preferences for different music mixes. A second result is that the model performance with values below 80% is still limited. On the one hand, this can be expected given the simplicity of the chosen approach (a fixed predefined feature set, straight-forward application of a standard pattern recognition method) and the limitations of the data set (one music piece, 19 conditions, 206 data points). On the other hand, this clearly shows that the present work is indeed a first proof-of-concept study which requires further experiments.

Next to these general results, we discuss the benefit of the different feature types by investigating the performance of the different models.

The LDR features alone (Model $M2$) show the worst performance, and removing those features from the full model ($M6$) is similar to the full model ($M1$). Thus, the LDR features appear not to be very useful for predicting the paired comparisons – at least not for the current data set.

The SPEC features alone ($M3$) show slightly lower performance than the full model ($M1$), and removing those

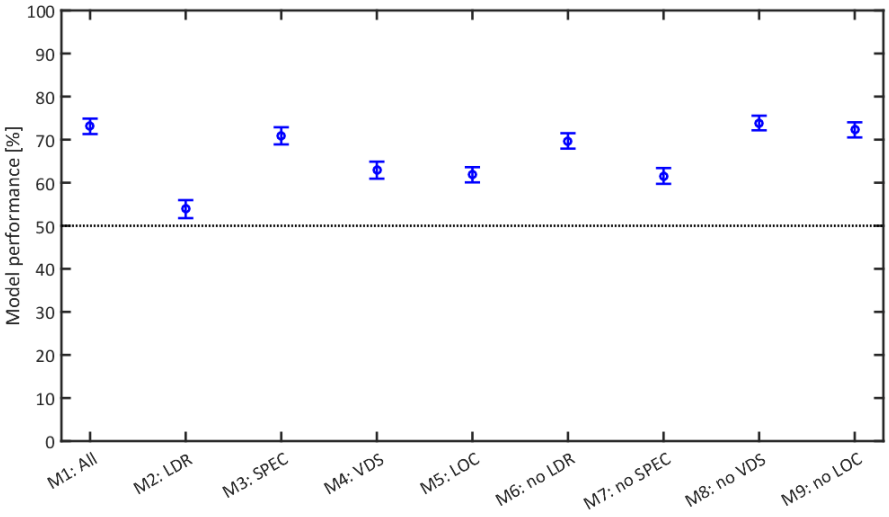

Figure 3: Model evaluation results for the nine different models. The plot shows the mean and 95% confidence intervals across 100 evaluation repetitions, using in each repetition a random data split of 80% training and 20% test data.

features from the full model ($M7$) causes a significant performance drop. Thus, the SPEC features appear to be very important for predicting the paired comparisons.

The VDS features alone ($M4$) show a significantly worse performance than the full model ($M1$), and removing those features from the full model ($M8$) has hardly any impact. Thus, the VDS features appear to have hardly any impact on the prediction performance. Essentially the same results can be found for the LOC features ($M5$ significantly lower than $M1$, hardly any difference between $M1$ and $M9$).

These results show that spectral features are required, but can be slightly supported by the other features. Interestingly, the other features appear to mutually cover their contributions to the model performance, a kind of redundancy that might be explained by the fact that those features address different spatial characteristics in a wider sense: two of the corresponding mixing parameters, *Reverb* and *Positioning*, are obviously related with spatiousness of individual sound sources, and also *Compression* might correlate with spatiousness as the modulation depth of the signals is changed.

Conclusions and future work

This study showed that it is in principle possible to predict human preference judgments for different mixing parameter settings of the same music piece. The results encourage to further investigate more advanced modeling approaches using more data, aiming for higher and more robust prediction performance. Another interesting direction of future work, however, is raised by the observation that several stimuli pairs do not show a clear preference ($PrefMat(m,n) \approx PrefMat(n,m)$). This suggests to further investigate the interaction of model performance and the characteristics of the ground-truth data, but requires more data to avoid overtraining effects as we have observed with our data set when running such analysis.

References

[1] ITU-R, "Recommendation BS.1116: Methods for the subjective assessment of small impairments in audio systems", International Telecommunications Union, Geneva, 2015.

[2] ITU-R, "Recommendation BS.1534: Method for the subjective assessment of intermediate quality levels of coding systems", International Telecommunications Union, Geneva, 2015.

[3] ITU-T, "Recommendation P.800: Methods for subjective determination of transmission quality", International Telecommunications Union, Geneva, 1996.

[4] ITU-T, "Recommendation P.805: Subjective evaluation of conversational quality", International Telecommunications Union, Geneva, 2007.

[5] C. Hold, H. Wierstorf, A. Raake, "Popmusik und Wellenfeldsynthese: Einfluss der Tonmischung", DAGA 2017.

[6] E. Skovenborg, "Measures of Microdynamics", 137th Convention of Audio Engineering Society, 2014.

[7] Two!Ears Consortium, "Auditory front-end", URL: `http://docs.twoears.eu/en/latest/afe/`, retrieved on Nov. 30, 2016.

[8] L. Nagel, "Predicting preference in productions of popular music with an auditory model", Master Thesis, Technische Universität Berlin, 2016.

[9] J. van Dorp Schuitman, D. de Vries, A. Lindau, "Deriving content-specific measures of room acoustic perception using a binaural non-linear auditory model", JASA Vol. 133, pp 1572–1585, 2013.

[10] International Society for Music Information Retrieval, "Overview of conference proceedings since 2000", URL: `http://www.ismir.net/conferences.html`, retrieved on Nov. 30, 2016.

[11] B. Efron, "Bootstrap Methods: Another Look at the Jackknife", The Annals of Statistics 7.1, pp. 1–26, 1979.

[12] J. Skowronek, M. McKinney, S. van de Par, "A demonstrator for automatic music mood estimation", 8th Intern. Conf. on Music Information Retrieval (ISMIR), 2007.

[13] C. Champagne, H. McNairn, B. Daneshfar, J. Shang, "A bootstrap method for assessing classification accuracy and confidence for agricultural land use mapping in Canada". International Journal of Applied Earth Observation and Geoinformation 29, pp. 44–52, 2014.