



IEA Heat Pumping Technologies



Acoustic Signatures of Heat Pumps

IEA HPT

Annex **51**

6.1: Annoyance rating and psychoacoustical analysis of heat pump sound

Henrik Hellgren, RISE, Sweden

Christian Kaseß, ARI, ÖAW Austrian Academy of Sciences, Austria

Date: November 2020	Review: 01
---------------------	------------



Index

1	Introduction	4
2	Annoyance rating of air source heat pump sound	4
2.1	Listeners	4
2.1.1	Austria	4
2.1.2	Sweden.....	4
2.2	Recording procedure	4
2.3	Annoyance rating.....	5
2.4	Data analysis.....	6
2.4.1	Psychoacoustic and acoustic parameters	6
2.4.2	Preprocessing.....	6
2.4.3	Consistency of the ratings.....	7
2.4.4	Statistical analysis.....	7
2.5	Results	8
2.5.1	Psychoacoustical and acoustical quantities	8
2.5.2	Annoyance	9
2.5.3	Annoyance index	12
2.6	Summary.....	12
3	Multidimensional scaling of experiences from geothermal heat pumps.....	13
3.1	Listeners	14
3.2	Recording procedure	14
3.3	Dissimilarity and preference rating	14
3.4	Results	14
3.5	Discussion.....	15
4	FIGURES INDEX.....	17



5	TABLES INDEX	18
6	REFERENCES	19



1 Introduction

Reduction of acoustic emissions is important to further increase the acceptance of heat pumps. To minimize the level of annoyance related to the noise experience it is important to understand which acoustic parameters influence the annoyance levels. Further, it is important that the methods to report noise levels correspond to the perceived annoyance.

2 Annoyance rating of air source heat pump sound

2.1 Listeners

2.1.1 Austria

20 normal hearing listeners (10 female) were tested. The mean age was 29.7 ± 6.8 years. All but one listener had hearing thresholds less than 20 dB higher than normal thresholds for all frequencies tested. A single listener had a single sided increase in hearing threshold of 30 dB at 8000 Hz but had otherwise normal hearing.

2.1.2 Sweden

/Results related to the responses of the Swedish listeners will be updated. Experiments delayed due to the Covid-19 situation /

10 normal hearing listeners (2 female). The mean age was 47.0 ± 10.4 years.

2.2 Recording procedure

The recordings were made in an hemi-anechoic room. The unit was an air-to-air heat pump with a heating capacity of approx. 6 kW at nominal condition. The recordings were made with free field microphones and a sampling frequency 51 200 Hz. The operation was controlled by adjusting the setting of the indoor unit fan speed. The recordings were made at five different operating conditions summarized by Table 2-1. The sound power level was determined according to ISO 3744. Each operating condition was recorded simultaneously at four microphone positions. The position at the right side was closest to the location of the compressor. The distance between the microphone and the unit was 1 meter. The microphone setup is shown in Figure 2-1. The recordings were 30 seconds long from which 5-second long sound samples were extracted to be used in the experiment.

Setting	Compressor speed [Hz]	Fan speed [rpm]	Input power [kW]	A-weighted sound power level, L_{WA} [dB]
Low	34	610	0.78	52.7
Medium	48	770	1.09	56.5
High	73	770	1.76	59.1
Super high	79	770	1.9	58.2
Emergency *	58	770	1.28	57.6

Table 2-1: List of recorded heat pump settings including fan and compressor speed and the measured A-weighted sound power level (according to ISO 3744).

* Emergency setting is a pre-defined program for test operation

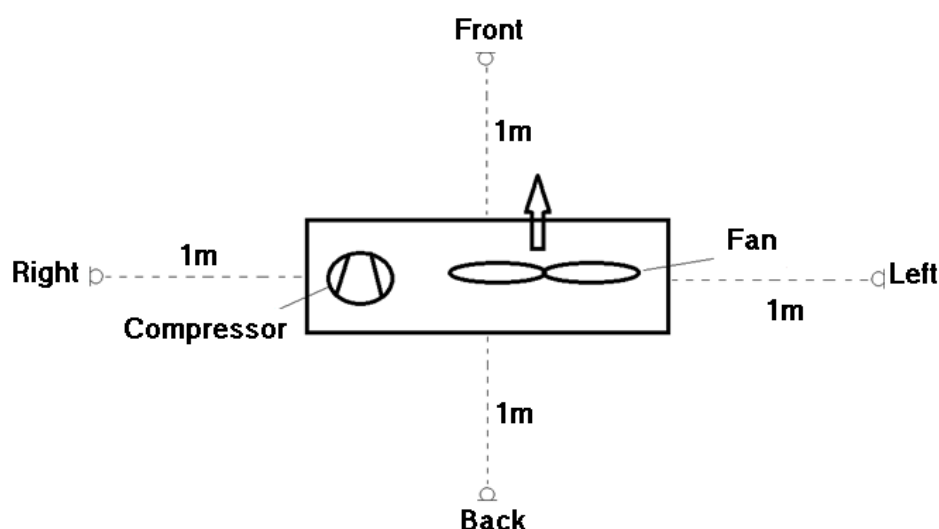


Figure 2-1: Microphone setup for acoustic measurements.

2.3 Annoyance rating

A free magnitude estimation was performed to determine the annoyance ratings of the different heat pump noises [1] [2], thus the listeners judged the relative annoyance rather than an absolute impression which is highly context dependent [3]. After listening to the stimulus, listeners were asked to input a numerical rating corresponding to the perceived annoyance. While listeners were free in choosing their starting value, they were instructed to avoid extremely high or low starting values in order to stay within a comfortable range of numbers. Listeners were asked to perform a proportional rating, i.e. double the annoyance should result in doubling the value.

Listeners were also instructed not to use 0 or negative numbers. They were also explicitly told to keep their rating scale constant within and across all runs. Once the rating was entered, listeners continued by pressing a key.

Before the main test, subjects received written instructions containing the definition of annoyance and a description of the procedure in the respective language. For this the instructions were first derived in English and then translated into German and Swedish.

Annoyance was defined as a feeling of discomfort, caused by noise or a feeling of aversion, discomfort, or irritation if the current activity is disturbed or affected by noise. Subjects were also asked to base their annoyance rating on imagining how annoying and distracting they would find the noise, if they were subjected to it on a regular basis. After reading the instructions, listeners performed a training covering a range of stimuli. The training consisted of a few trials, after which listeners were allowed to adapt their rating range in the case they felt uncomfortable with their initial choice. After the training, subjects had the opportunity to clarify open issues.

The experiment was performed in three runs in which each stimulus appeared three times. For each listener and run, stimuli order was randomized. Between runs a break of at least 5 minutes was enforced.

2.4 Data analysis

2.4.1 Psychoacoustic and acoustic parameters

Acoustic as well as psycho-acoustic parameters of the 5-second long sound samples were calculated using the Matlab-toolbox pysound3 [4]. These quantities encompassed loudness based on the Glasberg und Moore model [5], psychoacoustic roughness [6], tonality [7], sharpness, and loudness fluctuation [8]. Furthermore, C-weighted sound pressure levels (time-weighting fast) were calculated. The median as well as the 5%-percentile (the value that is exceeded 5% of the time) were calculated, denoted e.g. as S50 and S5 for the median and the 5% sharpness. The loudness level in phon was also determined. For all segments, the first 500 ms were discarded to avoid systematic errors due to transient response of the models.

2.4.2 Preprocessing

Three subjects reported a total of four input typos all of which were reproducible and could be corrected. As the magnitude estimation leads to a ratio scale, we applied the logarithm of base 2 on the data. Thus an increase by 1 in the log-ratings implies a doubling of the perceived annoyance. No were detected outside the 3-fold standard deviation across the subject data. The overall consistency of the ratings was good.

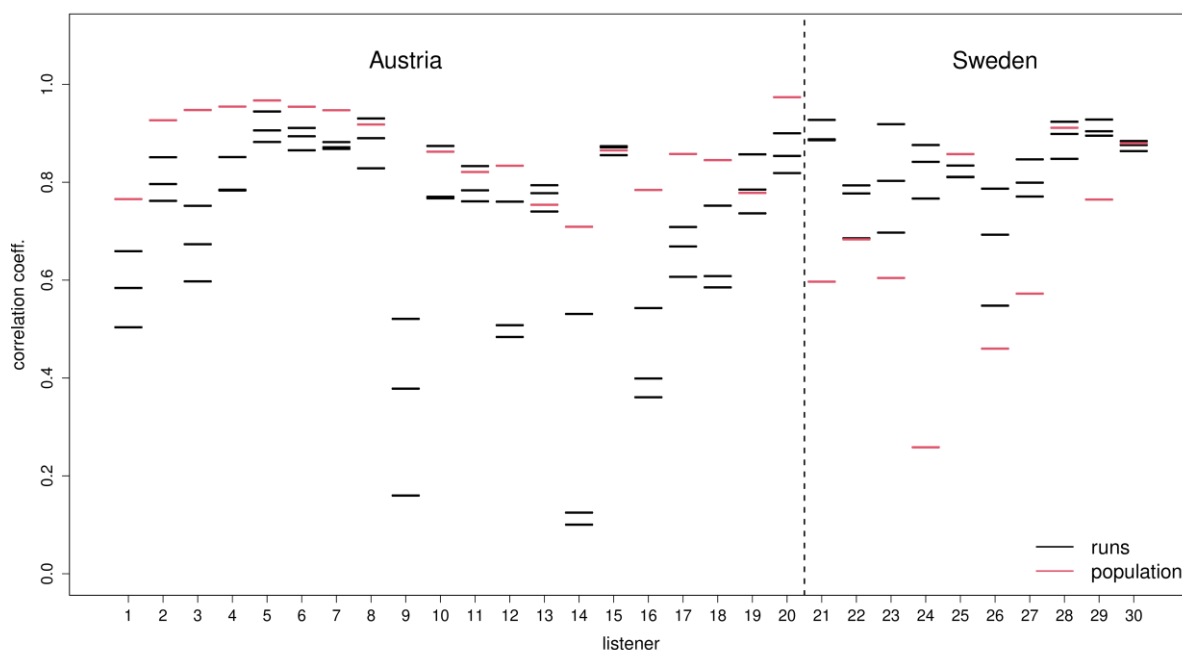


Figure 2-2: Correlation coefficient per listener between runs (black) and for the average population rating (red)

2.4.3 Consistency of the ratings

Figure 2-2 shows the correlation between the runs per listener (3 combinations, black symbols) as well as the correlation of the average response per listener and the population response (red symbols). It is clear that most listeners were able to produce consistent ratings across runs. Notably, at least two listeners seemed to have some difficulties exhibiting very low correlations. For listener 14 the ratings in the third run did not correspond to the first and second run which may be a sign of fatigue. The mean rating was still somewhat related to the group average. Listener 9 also showed low between-run correlations and a low negative correlation to the population. Compared to the overall mean, the Swedish listeners produced consistent results with lower overall correlation to the grand mean which is dominated by the Austrian listeners due to the higher number of listeners. When comparing the grand mean over all subjects from Sweden and Austria, a high correlation of 0.96 was observed.

For further analysis, mean log ratings per listener and condition were calculated and the grand mean per listener was subtracted in order to normalize the data. (see [1] [2]). For the group mean the average across all listeners were calculated per condition.

2.4.4 Statistical analysis

Statistical analysis was performed using the software R [9]. The mean log-ratings per listener and condition were the input for a repeated-measures-analysis-of variance (RM-ANOVA) with operating condition and direction as factors. The R-package afex was used for this purpose [10]. For significant effects omnibus post-hoc tests were performed using multivariate testing using emmeans (Estimated Marginal Means (Least-Squares Means)) [11].

Furthermore, the relation between acoustical properties and the annoyance rating was investigated using a stepwise linear regression. For this the function stepAIC (R-package MASS) was used which allows for both, adding and removing parameters. To determine the model quality, the Bayes Information Criterion (BIC) was used that allows us to take into account model fit and complexity [12].

2.5 Results

2.5.1 Psychoacoustical and acoustical quantities

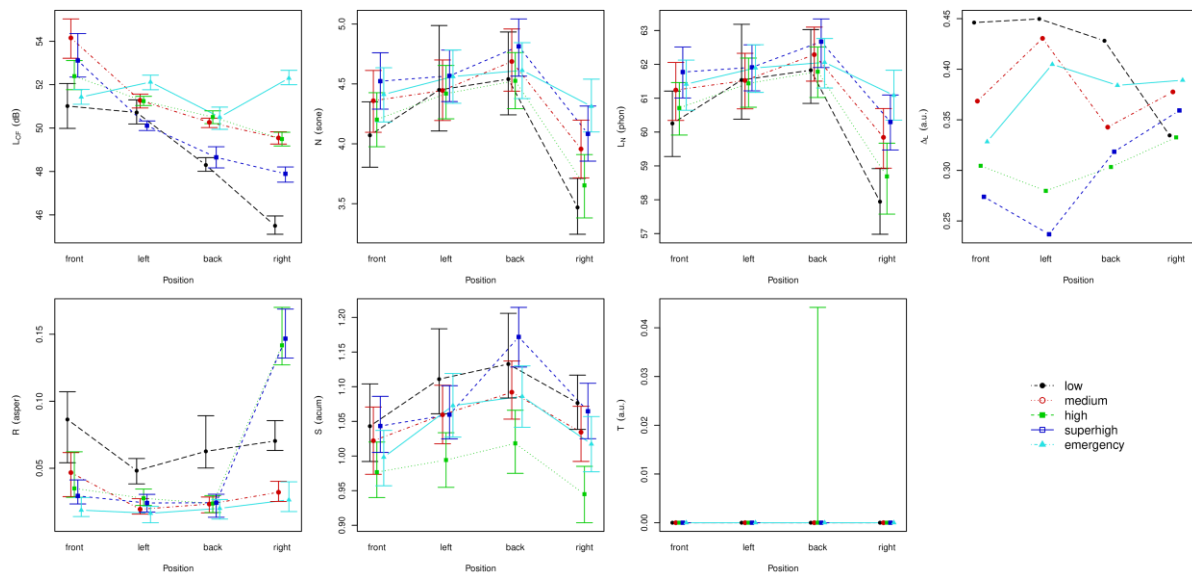


Figure 2-3: Acoustic descriptors as a function of position. Operating condition is shown as different colors.

Figure 2-3 shows the median and inter-quartile-range of the different acoustical quantities over time as a function of position after equalization to 40 dB(A). For the L_C and to a lower degree also the loudness (N) and loudness level (L_N) the different compressor speeds show most clearly up in the right position, where the compressor was located. There is also a visible effect of position for these two quantities, whereas for sharpness (S) the fluctuations are in the range of the effect. Roughness (R) is slightly elevated for the “low” condition and in the right position high and superhigh settings produce elevated roughness. For tonality only two conditions lead to non-zero peak values and only one condition produced non-zero values for at least a quarter of the time. Figure 2-4 shows the same data as Figure 2-3 arranged as a function of operation condition.

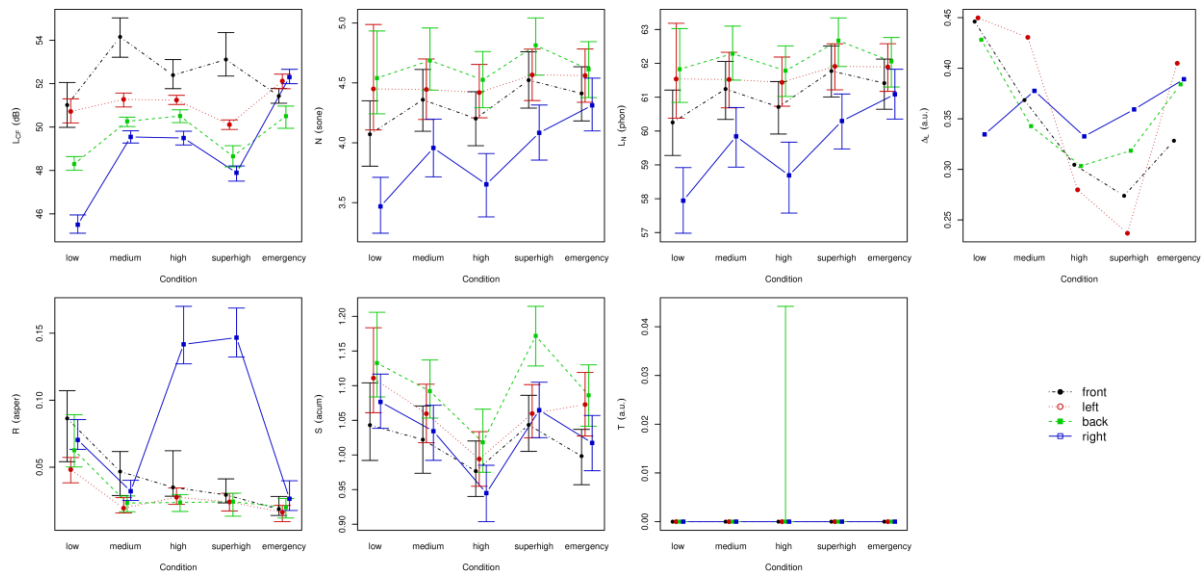


Figure 2-4: Acoustic descriptors as a function of operating condition. Recording position is shown as different colors.

2.5.2 Annoyance

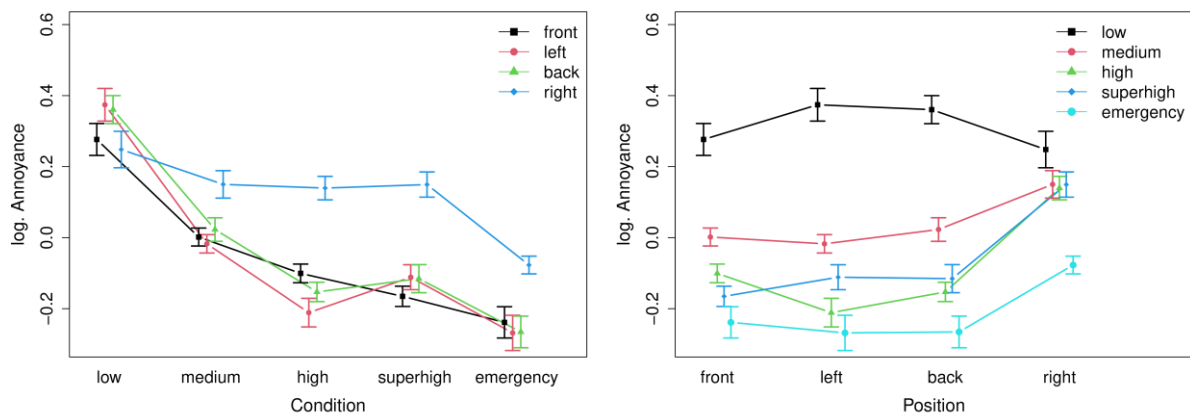


Figure 2-5: Mean and standard error of the annoyance as a function of operating condition and position

Figure 2-5 shows mean and standard error of the annoyance ratings across the Austrian study population. The ANOVA yielded significant main effects of position and condition as well as a significant interaction between the two factors ($p < 0.0001$ for all effects). Mauchly's test for sphericity showed a significant deviation from the sphericity assumption, thus a Greenhouse-Geisser correction was applied [13]. After correction, all effects were still significant with $p_{GG} < 0.0001$.

For four levels in position and five levels in condition a total of 60 possible pairwise interactions exist for which a post-hoc analysis was performed. P-values were Bonferroni-corrected, i.e. with the number of post-hoc test performed. The main results of this analysis is that all 16 significant interactions include either the right position or the low condition or both.

Thus, main effects containing either of these levels have to be treated with caution.

2.5.2.1 Main effect position

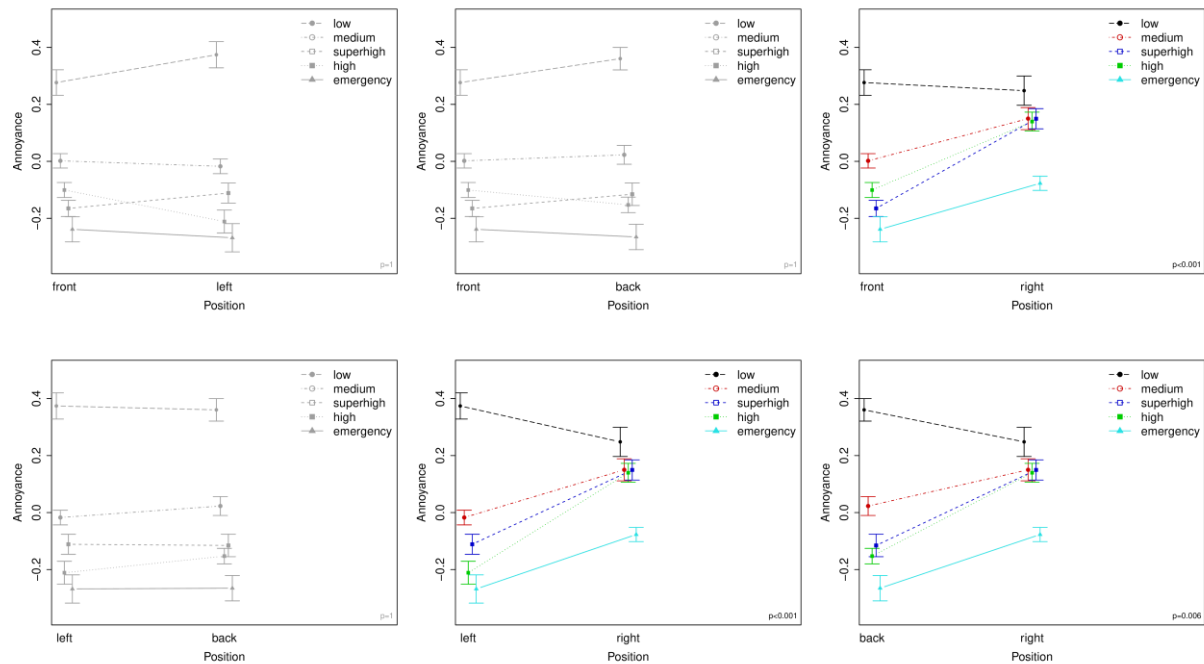


Figure 2-6: Pairwise post-hoc results for the main effect position. Gray plots are not significant after correction ($p > 0.05$).

A post-hoc test on all possible main effects between 2 positions shows, that the recording from the right position is significantly more annoying than all other positions. However, clearly when looking at the different contributions of the condition (Figure 2-6), the low condition has the opposite effect which is also significant for all but one pairwise interactions between the respective positions and the remaining conditions (i.e. front-right vs low-emergency). Due to this significant qualitative interaction effect the position effect cannot be properly interpreted as such.

2.5.2.2 Main effect conditions

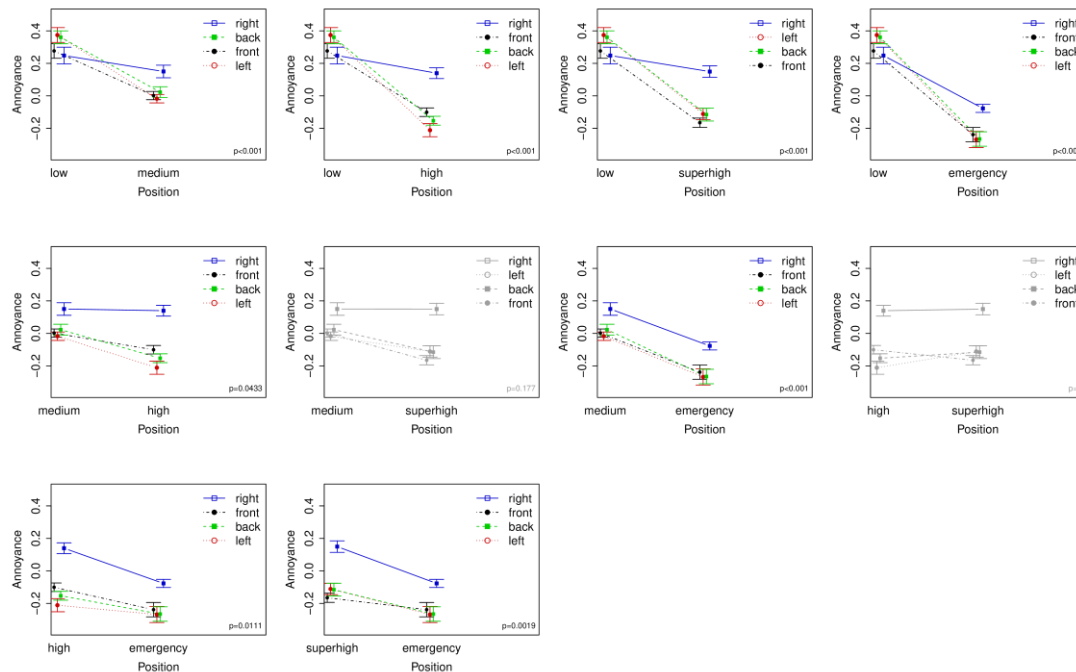


Figure 2-7: Pairwise post-hoc results for the main effect condition. Gray plots are not significant after correction ($p > 0.05$).

Figure 2-7 illustrates the post-hoc results for the main effect condition. Most conditions are significantly different from each other with the exception of superhigh to medium and high. Although some significant interaction also effect significant post-hoc contrasts, all these interactions are of a quantitative nature, i.e. they do not alter the direction of the effect. For example, for low-medium the right position shows an interaction with all other positions, however still the low condition is more annoying than the medium condition for all positions. A borderline effect is medium vs high were in the right position the annoyance is virtually constant.

2.5.3 Annoyance index

The statistical analysis showed significant differences between the conditions and positions. Using a stepwise model selection scheme an annoyance index is derived based on different acoustic descriptors Figure 2-8.

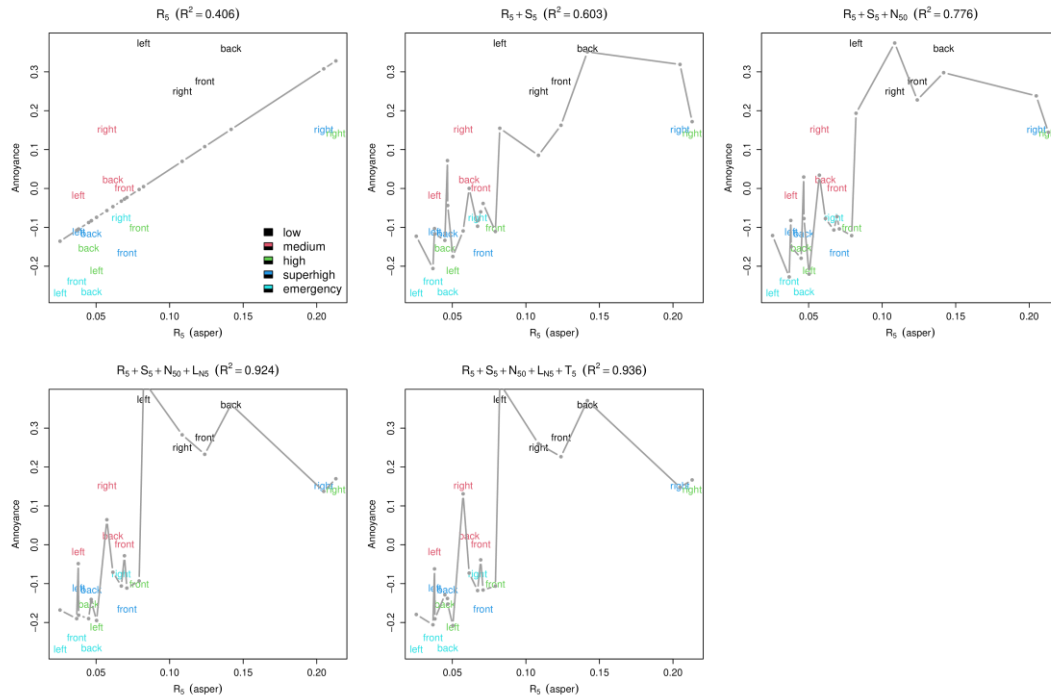


Figure 2-8: Stages of the step-wise model selection. The response is plotted vs. the single best descriptor.

The best model consisted of the following 5 descriptors:

$$\text{log annoyance} = \text{const} + 0.8891 \cdot R_5 + 1.2448 \cdot S_5 - 1.1498 \cdot N_{50} + 0.2512 \cdot L_{N5} + 1.1501 \cdot T_5$$

The single best descriptor is the peak psychoacoustic roughness which explained about 40% of the variance. Peak sharpness, median loudness and peak loudness level explain roughly an additional 20%, 15% and 15%, respectively. Peak tonality had only a minor effect and no other descriptor would have produced any gain at this stage of the stepwise model selection process.

2.6 Summary

Summarizing, from the current data a number of conclusions based on the data from the Austrian site can be drawn:

A main effect of the operating condition on the annoyance was observed. In particular the low-compressor speed condition was judged the most annoying whereas the emergency condition was judged less annoying than any other condition. In between the effects were minor. From

the annoyance index the low annoyance of the emergency mode could be explained by a low psychoacoustic roughness. However, for the low-condition roughness as well as sharpness do not seem to be the main contributing factor. Interestingly, adding loudness seems to improve in particular on the fit for this condition, however, the tendency is reversed such that lower median loudness increases annoyance at this stage.

Position does not lead to a consistent main effect since interactions between position and condition are present. In particular the low-condition all but the right measurement position produce comparatively high annoyance ratings and thus result in this interaction effect.

A preliminary comparison of the two sites where the experiment took place indicates that the tendencies in the judgments are relatively similar. As only half the planned listeners have been tested, a detailed analysis will be performed once sufficient data is available.

3 Multidimensional scaling of experiences from geothermal heat pumps

To determine the most salient parameters influencing perception of geothermal heat pumps and the corresponding level of annoyance a dissimilarity rating was conducted along with a preference mapping. Dissimilarity ratings are powerful tools to obtain a multidimensional scaling of the stimuli, free of the restrictions imposed by predetermined scales or response criteria. It builds on the limited ability of the listener to only focus on a set of varying parameters [14]. To determine the prevalent or dominant perceptual features in different geothermal heat pumps the dissimilarity rating conducted included three different models and 10 different recordings of varying situations (3+3+4 of the three models). The corresponding multidimensional map was compared with specific psychoacoustic parameters as well as rated level of annoyance.

Few studies have been conducted on the perception and experience from geothermal heat pumps. A study by Persson Waye and Rylander compared heat pumps and ventilation systems dominated by lower frequencies (<200Hz) and heat pumps and ventilation systems dominated by mid frequencies [15]. The results showed that people exposed to low frequency noise from heat pumps were more annoyed and had a higher level of disturbed concentration than those exposed to the noise of mid-frequency character. Wang and Novak analysed several different heating, ventilation and air-conditioning systems, they determined that high sound levels (>50 dBA), excessive low frequency rumble and larger timescale fluctuations (e.g., a heat pump cycling on and off every 30 seconds) were the dominating characteristics influencing levels of annoyance [16].

Annoyance to mechanical systems related to heating appear to often be related to the dominance of low frequency content. Broner and Leventhall proposed using the difference between A-weighted SPL and C-weighted SPL that values greater than 20 dB would signify a low frequency noise problem [17]. Holmberg et al suggested that the problem would occur already at 15 dB. In the present listening test three stimuli had a greater difference than 15 dB (a1, b1, and c2) whereof one had a greater difference than 20 dB (c2) [18].

3.1 Listeners

In the listening test 14 people participated, 4 women and 10 men (M= 40 years old, s.d. = 9 years). 1 participant did not comply with the instructions and was removed from further analysis. 1 participant reported hearing problems, but that did not affect the results.

3.2 Recording procedure

Three different geothermal heat pumps were used in the experiment. Each heat pump was represented by three or four different recordings. In total 10 stimuli were utilized. All stimuli were 3 seconds long and presented at 30 dB(A). The sound pressure level choice was made as the current labelling is done using dB(A) levels instead of loudness measures.

Model	C-weighted level (dB(C))	Loudness (sone)	Roughness (asper)	Sharpness (acum)	Compressor speed (Hz)	Rated preference
a1	60.21	3.78	0.066	1.85	45	3.9
a2	55.78	3.63	0.125	1.51	50	2.8
a3	48.49	4.44	0.056	1.71	57	6.3
b1	61.85	3.65	0.018	1.25	41	4.9
b2	53.73	3.77	0.071	1.47	52	5.4
b3	40.71	4.14	0.030	1.33	110	7.0
c1	57.38	3.85	0.096	1.17	58	3.6
c2	63.49	3.92	0.062	0.88	69	6.5
c3	56.75	3.78	0.080	0.88	83	4.8
c4	48.87	3.73	0.094	1.19	100	6.3

Table 3-1: Psychoacoustical properties and settings for the different heat pumps.

3.3 Dissimilarity and preference rating

The listening test took place in a 3rd order ambisonics lab with little visual distraction. The sounds presented were mono sounds presented using the two front speakers. Each participant performed pairwise ratings of dissimilarities between the different sounds using a sliding scale. In addition the participants marked which of the sound in the pair s/he preferred. Using a half-matrix design (testing all possible pairs in one direction) this resulted in 45 pairs. To be noted: the participants were aware of the sounds coming from different geothermal heat pump systems as it could affect their choice of preference.

3.4 Results

The dissimilarity ratings were analysed using the individual difference scaling (INDSCAL) model. The INDSCAL model assumes that all participants share the same psychological scale but attends differently to the underlying psychological dimensions (Ashby et al, 1994). An advantage with the INDSCAL model is that it provides a unique configuration solution that requires no further rotation of the model (Martens and Zacharov, 2000). The analysis resulted in a 2-dimensional model (Stress=.131). Stress values <.133 were considered acceptable as determined by Sturrock and Rocha (2000). The MDS solution is presented in Figure 3-1 labelled by their model (a-c).

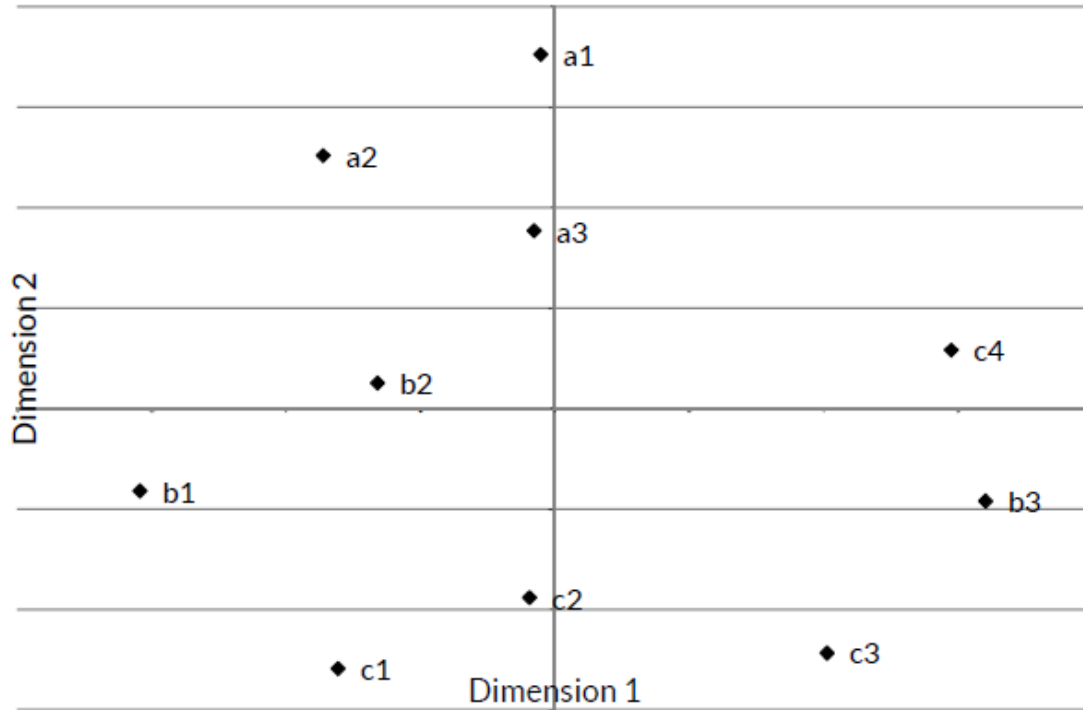


Figure 3-1: The MDS solution labelled by their brand.

The two dimensions were analysed using the preference ratings of the listening test, the psychoacoustic parameters and the compressor speed. The preference ratings are listed in Table 3-1. The results showed that Dimension 1 is partly explained by the preference mapping ($R^2 \text{ adj.} = .32$, $F = 5.2$, $p < .05$) but mostly by the compressor speed ($R^2 \text{ adj.} = .83$, $F = 46.2$, $p < .001$). Dimension 2 is explained by the variance in sharpness ($R^2 \text{ adj.} = .74$, $F = 27.0$, $p < .001$). The other psychoacoustic parameters showed no significant relationship with either dimension. Regression analyses further showed that the preference mapping could be explained by both compressor speed and sharpness ($R^2 \text{ adj.} = .67$, $F = 10.0$, $p < .01$), the participants preferred sounds with less sharpness and a compressor speed at higher frequency.

3.5 Discussion

Little of previous research on the sounds of heat pumps have focused on other aspects than low frequency content and tonality. This experiment is a first step to further distinguish the dominating parameters to explain perception of ground source heat pumps. Creation of perceptual maps require an inter comparison between the specific stimuli used in the experiment. The result will thus depend on which stimuli are used. The aim of the experiment was to use as different heat pump sounds as possible to set a ground work for later experiments on finding the parameters explaining annoyance for heat pumps. The experiment was limited to ground source heat pumps, as we believe that air source heat pumps has a distinct different sounds, the latter will instead be evaluated in a later experiment.

The low frequency content did not influence the level of annoyance. This might seem surprising, but Kjellberg et al (1997) proposed that the difference between the C-weighted and the A-weighted SPL may be limited as predictor of annoyance when the overall noise level is too low. This could be a reason to the lack of connection between annoyance and dB(C) in



present study. However, most ground source heat pumps hold a relatively low sound pressure level, indicating that dominating low frequency character might not influence the annoyance level to a higher degree.

Dissimilarity ratings require the use of shorter sound stimuli to enable comparison between presented pairs. This makes it difficult to discern whether fluctuations in the heat pumps influence level of annoyance. For future experiments longer stimuli are needed.

The results showed that the most salient parameters are compressor speed and the sharpness level. Both have a significant impact on annoyance responses to the ground source heat pumps. To further evaluate whether fluctuations also influence annoyance longer stimuli are needed.



4 FIGURES INDEX

Figure 2-1: Microphone setup for acoustic measurements.	5
Figure 2-2: Correlation coefficient per listener between runs (black) and for the average population rating (red).....	7
Figure 2-3: Acoustic descriptors as a function of position. Operating condition is shown as different colors	8
Figure 2-4: Acoustic descriptors as a function of operating condition. Recording position is shown as different colors.....	9
Figure 2-5: Mean and standard error of the annoyance as a function of operating condition and position.....	9
Figure 2-6: Pairwise post-hoc results for the main effect position. Gray plots are not significant after correction ($p>0.05$).	10
Figure 2-7: Pairwise post-hoc results for the main effect condition. Gray plots are not significant after correction ($p>0.05$).	11
Figure 2-8: Stages of the step-wise model selection. The response is plotted vs. the single best descriptor.....	12
Figure 3-1: The MDS solution labelled by their brand.	15



5 TABLES INDEX

Table 2-1: List of recorded heat pump settings including fan and compressor speed and the measured A-weighted sound power level (according to ISO 3744).	5
Table 3-1: Psychoacoustical properties and settings for the different heat pumps.	14



6 REFERENCES

- [1] M. Nilsson, M. Andehn and P. Lesna, "Evaluating roadside noise barriers using an annoyance-reduction criterion," *J. Acoust. Soc. Am.*, vol. 124, pp. 3561-3567, 2008.
- [2] C. Kasess, P. Majdak, T. Maly and H. Waubke, "The relation between psychoacoustical factors and annoyance under different noise reduction conditions for railway noise," *J. Acoust. Soc. Am.*, vol. 141, pp. 3151-3163, 2017.
- [3] S. Fidell, "Noise-Induced Annoyance," in *Handbook of Noise and Vibration Control*, Hoboken, New Jersey, John Wiley and Sons, 2007, pp. 316-319.
- [4] D. Cabrera, S. Ferguson, F. Rizwi and E. Schubert, "PsySound3: software for acoustical and psychoacoustical analysis of sound recordings," in *Proceedings of the 13th International Conference on Auditory Display*, Montreal, Canada, 2007.
- [5] B. Glasberg and B. Moore, "Derivation of Auditory Filter Shapes from Notched Noise Data," *Hearing Research*, vol. 47, pp. 103-137, 1990.
- [6] P. Daniel and R. Weber, "Psychoacoustical roughness: implementation of an optimized model," *Acustica*, vol. 83, pp. 113-123, 1997.
- [7] E. Terhardt, G. Stoll and M. Seewan, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *J. Acoust. Soc. Am.*, vol. 71, pp. 679-688, 1982.
- [8] J. Chalupper and H. Fastl, "Dynamic Loudness Model (DLM) for Normal and Hearing-Impaired Listeners," *Acta Acustica United with Acustica*, vol. 88, pp. 378-386, 2002.
- [9] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, 2018.
- [10] H. Singmann, B. Bolker, J. Westfall, F. Aust and M. S. Ben-Shachar, "afex: Analysis of Factorial Experiments," 2020.
- [11] R. Lenth, "emmeans (Estimated Marginal Means (Least-Squares Means))," 2020.
- [12] W. Venables and B. Ripley, *Modern Applied Statistics with S* Fourth Edition, New York: Springer, 2002.
- [13] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461-464, 1978.
- [14] J. R. Miller and E. C. Carterette, "Perceptual space for musical structures," *J. Acoust. Soc. Am.*, vol. 58, pp. 711-720, 1975.
- [15] K. Person Waye and R. Rylander, "The prevalence of annoyance and effects after long-term exposure to low-frequency noise," *Journal of Sound and Vibration*, vol. 240, pp. 483-497, 2001.
- [16] L. M. Wang and C. C. Novak, "Human Performance and perception-based evaluations of indoor noise criteria for rating mechanical system noise with time-varying fluctuations," *ASHRAE Transactions*, vol. 116, pp. 553-568, 2010.



- [17] N. Broner and H. G. Leventhall, "Low frequency noise annoyance assessment by low frequency noise rating (LFNR) curves," *Journal of Low Frequency Noise, Vibration and Active Control*, vol. 2, pp. 20-28, 1983.
- [18] K. Holmberg, U. Landström and A. Kjellberg, "Low frequency noise level variations and annoyance in working environments," *Journal of Low Frequency Noise, Vibration and Active Control*, vol. 16, pp. 81-87, 1997.