inter.noise 2004
Prague • Czech Republic • August 22–25

The 33rd International Congress and Exposition
on Noise Control Engineering

Predicting annoyance judgments from psychoacoustic metrics: Identifiable versus neutralized sounds

W. Ellermeier[a], A. Zeitler[a] and H. Fastl[b]

[a]Sound Quality Research Unit, Department of Acoustics, Aalborg University,
Fredrik Bajers Vej 7 B-5, DK-9220 Aalborg Ø, Denmark
[b]AG Technische Akustik, Mensch-Maschine-Kommunikation, TU München, Germany

[a][we;az]@acoustics.dk;[b]fastl@mmk.ei.tum.de

Abstract [267] It is common practice to predict perceived noise annoyance by means of regression models using instrumental psychoacoustic metrics as predictors. The validity of this approach has been criticized for not taking into account non-sensory variables such as the meaning of the sound. The present study investigates to which extent judgments of annoyance reflect sensory attributes in terms of psychoacoustic metrics as opposed to cognitive and emotional variables related to the sound source. A new signal-processing method which substantially reduces the identifiability of sound sources was applied to a set of 40 environmental and product sounds. In the listening experiment, two independent groups of participants ($n = 25$ each) provided annoyance judgments of either the original or neutralized version of the sounds using the method of category-subdivision scaling. In the second part of the experiment, the participants rated the affective meaning of the sounds on a concept-specific semantic differential. Instrumental analyses of the sounds included the calculation of psychoacoustic metrics of loudness, sharpness, roughness, fluctuation strength, and tonal prominence which were entered into regression models to predict the outcome of the listening tests. It turned out that while instrumental metrics fared well in predicting overall annoyance, they did not account for the discrepancies in judgments of original versus neutralized sounds, suggesting that these actually reflect non-sensory effects mediated by the 'meaning' of the sound.

1 INTRODUCTION

Psychoacoustic metrics (indices of loudness, sharpness, roughness, and the like) when properly computed, and combined, usually go a long way in predicting the annoyance reactions produced by environmental sounds. They cannot, however, account for non-sensory influences entering into the judgment of a sound. These might include the attitude towards the source, effects of familiarity, preferences, and user expectations about prototypical sounds. Such non-auditory influences on psychoacoustical judgments are sometimes summarized as effects of the *meaning* of the sound [1, 2].

Due to Fastl's recent proposal of a signal-processing algorithm [3] that modifies the acoustic properties of a given sound so that it is very likely to become unrecognizable (and thus 'meaningless' in the sense discussed), we now have a handle on quantifying the relative contribution of acoustic and non-acoustic factors to the judgments made in a listening test. The advantage of Fastl's method over other alternatives (such as filling the temporal envelope of the original sound with broadband noise) is that it takes both temporal and spectral properties into account, and is designed to preserve the temporal loudness pattern of the original. That is accomplished by first subjecting the sound to a Fourier time transform (FTT), then applying some spectral broadening to the elements of the FTT pattern, and subsequently re-synthesizing the sound by an inverse FTT [3].

In an earlier report [4], we focused on the effects of this 'neutralization' procedure on *loudness* judgments, showing that while identifiability was greatly reduced (from 93 to 13%), the effects on loudness were relatively small, reaching significance only for three in a sample of 40 sounds. In the present report, we will analyze the effects of the neutralization procedure on ratings of *annoyance*. Clearly, while the procedure guarantees loudness (as measured by appropriate metrics) to be unchanged, when investigating annoyance ratings of original and neutralized sounds, both acoustical differences (changes in roughness, for example due to the spectral broadening, see [5]), and differences in 'meaning' will be confounded. To disentangle these effects, in a first step, annoyance ratings of both kinds of sounds will be obtained from two independent groups of subjects. Subsequently, the results will be related to (a) psychoacoustic metrics capturing the sensory effects of the sounds, and to (b) semantic-differential ratings capturing their connotative 'meaning.'

2 METHOD

2.1 Participants

A total of 50 students at Aalborg University between 19 and 31 years of age participated in the experiment. They were audiometrically screened with the requirement that their pure-tone thresholds did not exceed the normal curve by more than 20 dB in the frequency range from 0.25 to 8 kHz. Subsequently, half of the participants were randomly assigned to judge the annoyance of the original, half to judge the annoyance of the neutralized sounds.

2.2 Apparatus and Stimuli

The original sounds were recorded using a Brüel & Kjær (Portable PULSE 3560 C) frontend connected to a (mono) microphone (Brüel & Kjær type 4165 or 4179) placed at appropriate distances from 0.3 to 7 m from the source. The files were converted to 16-bit, 44.1 kHz format to be played from a regular (RME Digi96 Pro) sound card the output of which was amplified (Behringer HA 4400) before being presented diotically to the subjects listening in a double-walled sound-attenuating chamber via headphones (Beyerdynamic DT 990).

Fourty sounds were selected for the experiment to be highly identifiable in the original condition: Most of them were non-stationary everyday noises (e.g. toilet flush, door closing, scissors, car passing), about a third of the sounds consisted of product sounds of electrical devices (e.g. hairdryer, kitchen mixer, razor) recorded in their typical use. The sounds varied in duration from 0.7 to 5 s, and had overall sound-pressure levels between 30 and 80 dB SPL.

The 40 recorded sounds were processed using the algorithm proposed by Fastl [3] in order to obtain 40 "neutralized" sounds having identical loudness-time functions.

2.3 Procedure

All participants performed three tasks in the following order: (1) a scaling experiment (loudness or annoyance), (2) an identification task, and (3) a semantic-differential rating of all sounds. For the annoyance scaling task reported in this paper, a category subdivision procedure (CS, see [6]) was used: Subjects were asked to judge each sound on a combined verbal-numerical category scale that consisted of five verbal categories which were further subdivided into ten steps and labelled with the Danish equivalents of “very slightly annoying” (1-10), “slightly annoying” (11-20), “medium” (21-30), “strongly annoying” (31-40) and “very strongly annoying” (41-50). The endpoints of the resulting 50-point scale were verbally anchored to denote “not at all annoying” (0) and “unbearably annoying” (beyond 50). After a short practice run, each subject judged the sounds once in a random order. In the subsequent identification experiment, the 40 recorded (resp. neutralized) sounds were played again in a random sequence, and the subject was asked to identify the source by providing both a noun and a verb (e.g. “motor - idling”). During a second session, subjects judged the same sounds using a semantic differential consisting of 12 bipolar adjective scales.

3 RESULTS

In the present report we will focus on the annoyance scaling data, and their relation to instrumental psychoacoustic metrics. Preliminary results on loudness scaling, and on the outcome of the semantic differential measures have been reported elsewhere [4, 7].

3.1 Annoyance scaling

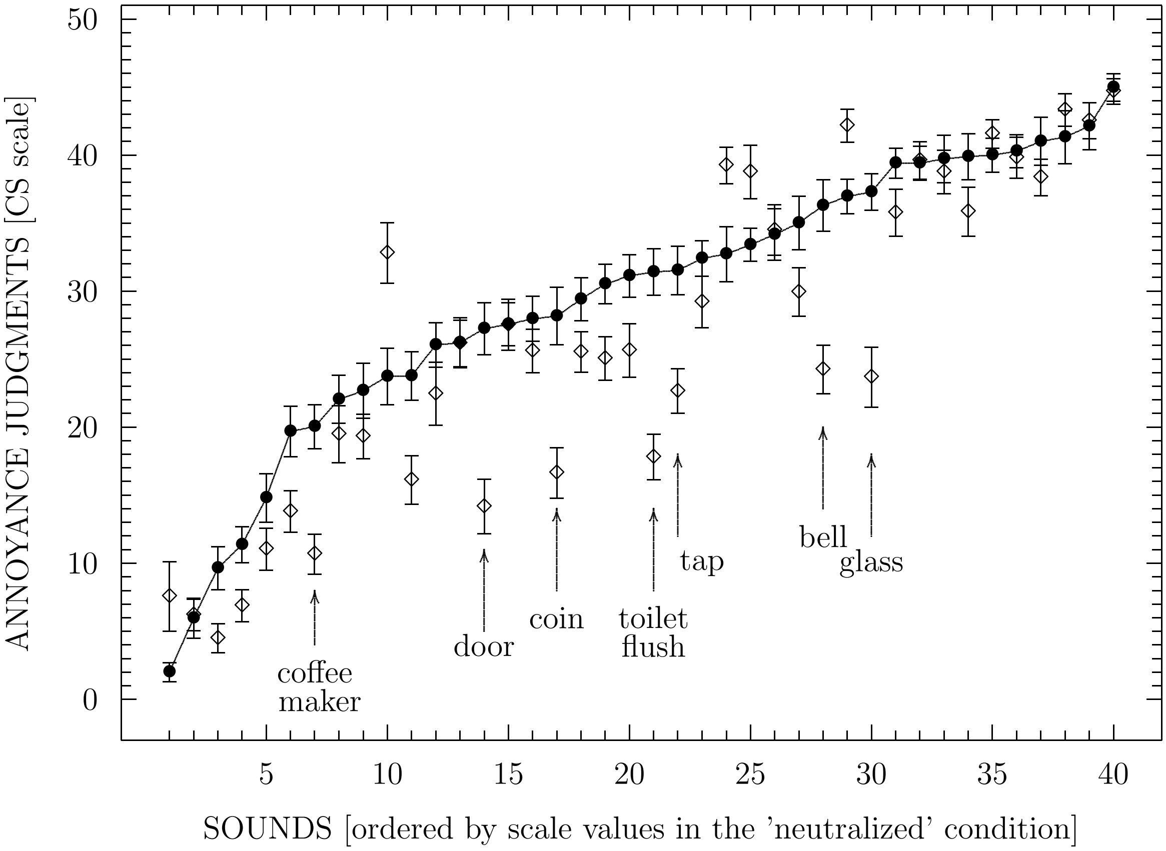

Figure 1: *Annoyance scale values (plus/minus standard errors of the means) of the 40 test sounds as judged by 25 participants in their neutralized (filled circles) and by a different sample of 25 participants in their original version (open diamonds). Statistically significant discrepancies are marked by vertical arrows.*

The annoyance scaling data were averaged across the 25 subjects in each group, and are displayed in Figure 1 with the sound samples (along the abscissa) being arranged in ascending order according to the mean annoyance produced by the neutralized sounds (filled circles), so that the judgments of the original, identifiable sounds (open diamonds) appear as deviations from the 'neutral' curve. It is evident that these deviations are substantial, which is confirmed by a two-factor (sound by processing), mixed analysis of variance: In addition to the highly significant (but trivial) main effect of the 40 sounds, there is a significant main effect of processing (original vs. neutral), indicating that annoyance ratings of the neutralized sounds were higher on the average than those of the originals: $F(1, 48) = 4.99$; $p < 0.03$. Furthermore, there is a highly significant (sound by processing) interaction, showing that the effects of the neutralization differ significantly between sounds: $F(16.25, 780.01) = 7.92$; $p < 0.001$.

In order to determine, which of the discrepancies between judgments of the original and neutralized sounds were producing these effects, post-hoc tests were performed using a Bonferroni correction to adjust for chance outcomes due to multiple testing. At the test-wise α-level of 0.00128 thus obtained, the seven sounds marked by arrows in Figure 1 were identified as producing significant differences in annoyance in the two experimental conditions investigated: coffee maker ($M_{neutr} - M_{orig} = 9.36$ scale units), door locking (13.08 scale units), bouncing coin (11.52), toilet flushing (13.6), water running from a tap (8.88), bicycle bell (12.04), champaign glass (13.6). Note that for all of these sounds annoyance judgments of the originals were lower than than those of the neutralized version.

3.2 Instrumental analyses

In the following, instrumental analyses of the 40 sounds and their neutralized counterparts were performed in order to explore whether changes in psychoacoustic metrics can account for the outcome of the annoyance scaling experiment. All analyses were performed using a commercially available psychoacoustic analysis program (Brüel & Kjaer Sound Quality type 7698, version 3.4.0).

Sound-quality changes due to the 'neutralization'

Since the parallel loudness scaling study of original and neutralized sounds [4] had shown no overall loudness differences due to processing ($M_{orig} = 28.62$; $M_{neutr} = 28.38$), we investigated whether other sound-quality metrics might be affected. As may be seen in Table 1, on the average the sounds slightly increased in fluctuation strength, and roughness, and decreased somewhat in sharpness. These observations are in close agreement with Fastl's [5] earlier analysis of a long-duration traffic noise recording. The most striking effect, however, is that - due to the spectral broadening - the sounds loose almost all of their tonal content (as indicated by the prominence ratio statistics in the last line of Table 1). In fact, for most of the neutralized sounds, no tonal components could be detected.

Analysis of the differences between original, and neutralized sounds

In a second step, psychoacoustic metrics were explored with respect to their potential in accounting for the differences between neutralized and original sounds (evident in Figure 1). To that effect, five sound quality metrics (the fifth percentile of non-stationary loudness N_5, and the four parameters listed in Table 1) were entered into a multiple-regression equation with the difference in annoyance ratings between neutralized and original sounds being the criterion (the value to be predicted). As a result, all sound-quality metrics, except for the prominence ratio dropped out as non-significant, and the latter accounted for less than 12 percent of the variance ($R^2_{adj} = 0.115$) in the differences to be predicted.

Table 1: *Changes in psychoacoustic metrics due to neutralization.*

	Stimuli	
	original	neutralized
Roughness [asper]	0.715	0.795
Fluctuation Str. [vacil]	2.070	2.315
Sharpness [acum]	3.090	2.875
Prominence ratio [dB]	1.985	-5.120

Note. Standard measures of roughness, fluctution strength, and the prominence ratio (substituting a 'threshold' value of -6 dB when no tonal component was detected) were computed for each sound, as well as the 50th percentile of Aures' sharpness. The table entries are median values of these metrics across the 40 sounds analysed.

Modelling overall annoyance

If psychoacoustic metrics do not do well in accounting for the *differences* in annoyance ratings due to processing, they might nevertheless be well suited in predicting overall annoyance in both kinds of sounds (originals, and neutralized ones). To explore this, the measured sound-quality indices of all 40 neutralized sounds were entered into a multiple linear regression. It turned out, that a combination of loudness (N_5), sharpness (S_{50}), and roughness (R) predicted the overall annoyance ratings (category scaling of annoyance, CSA) fairly well ($R^2_{adj} = 0.856$), accounting for nearly 86 percent of the variance:

$$CSA = 8.07 + 0.563 * N_5 + 3.022 * S_{50} + 2.175 * R \tag{1}$$

Using the same three metrics (while allowing for different coefficients) to predict the annoyance of the *original* sounds, reduced the variance accounted for by 15%: $R^2_{adj} = 0.704$. But even allowing for different metrics in the regression equation for the original sounds did not improve the situation: The best model (obtained when substituting the prominence ratio for sharpness) did only slightly better, $R^2_{adj} = 0.731$.

4 CONCLUSIONS

Direct scaling of the annoyance of well-recognizable sounds, and of their neutralized [3, 5] counterparts yields substantial differences. To explore, whether these differences are due to the acoustical changes inherent in the neutralization procedure, a number of analyses using instrumental sound quality metrics were performed, yielding three major conclusions: (1) The changes in psychoacoustical metrics due to the procedure are minor, except for a reduction in tonalness for those sounds having tonal components. (2) The differences in the directly scaled annoyance between original and neutralized sounds cannot be accounted for by any of the instrumental metrics explored. (3) Annoyance ratings of both the neutralized, and the original sounds, can be accounted for by fairly similar 'combination metrics,' but the model for the original sounds accounts for substantially less of the variance. The fact that the original sounds are less well predicted by psychoacoustic metrics suggests that the additional variance is caused by non-acoustic factors related to their identifiability.

These results suggest that the differences observed in the scaling of original vs. neutralized sounds are indeed due to differences in 'meaning,' i.e. to non-acoustical factors. This conclusion is also supported by preliminary analyses of the Semantic Differential ratings [7], supposedly measuring the 'connotative' associations elicited by the sounds. These ratings exhibit much higher correlations with the difference scores than do the instrumental metrics. Further analyses comparing both sets of predictors are under way.

ACKNOWLEDGEMENTS

This investigation was carried out while the first two authors were with the Sound Quality Research Unit (SQRU) at the Department of Acoustics, Aalborg University. This unit receives financial support from Bang & Olufsen, Brüel & Kjær, and Delta Acoustics & Vibration, as well as from the Danish National Agency for Industry and Trade (EFS), and the Danish Technical Research Council (STVF). The authors would like to thank Karin Zimmer for performing the regression analyses, and Signe Kristensen and Pelle Coltau for their conscientious assistance in helping with collecting and coding the data. Alfred Zeitler is now with the BMW sound engineering department: alfred.zeitler@bmw.de.

REFERENCES

[1] E. Zwicker, H. Fastl, Psychoacoustics. Facts and models, 2nd edition, Springer, Berlin 1999.

[2] U. Jekosch, Meaning in the context of sound quality assessment, *Acustica - acta acustica* **85(5)**, pp. 681–684, (1999).

[3] H. Fastl, "Neutralizing the meaning of sound for sound quality evaluations," in *Proceedings 17th International Congress of Acoustics (ICA 2001)*, Rome, Italy 2001, (CD-ROM).

[4] W. Ellermeier, A. Zeitler and H. Fastl, "Impact of source identifiability on perceived loudness," in *Proceedings of the 18th International Congress on Acoustics (ICA 2004)*, Kyoto, Japan, 2004, Vol. II, pp. 1491-1494.

[5] H. Fastl, "Features of neutralized sounds for long term evaluation," in *Proceedings Forum Acusticum 2002*, Sevilla, Spain, 2002, (CD-ROM).

[6] J. Hellbrück, "Category subdivision scaling - A powerful tool in audiometry and in noise assessment," In H. Fastl et al. (Eds.), *Recent trends in hearing research. Festschrift for Seiichiro Namba*, BIS, Oldenburg 1996, pp. 317-336.

[7] A. Zeitler, W. Ellermeier and H. Fastl, "Significance of meaning in sound quality evaluation," in *Proceedings SFA/DAGA 2004*, Strassbourg, France, 2004.