Impact of Source Identifiability on Perceived Loudness

Wolfgang Ellermeier, Alfred Zeitler*, Hugo Fastl***

* Sound Quality Research Unit (SQRU), Dep. of Acoustics, Univ. Aalborg, Denmark
** AG Technische Akustik, MMK, TU München, Germany
we@acoustics.dk

Abstract

In order to disentangle acoustical and non-acoustical factors in sound-quality evaluation, a signal-processing scheme [H. Fastl (2001). Proceedings 17th ICA, Rome] that renders most sounds unrecognizable while preserving their temporal loudness envelopes was employed. Two independent groups of subjects (N=20 each) evaluated a set of 40 environmental and product sounds, either in their original, or in a thus processed version. In contrast to previous investigations using this methodology, the stimuli covered a larger range of sound pressure levels, and additional data were collected on the proportion of correct identifications in the two versions presented. The results show that the identifiability is greatly reduced (from 90 to 13% on the average) by the "neutralization" procedure. Loudness judgments of the original and neutralized sounds are indistinguishable for the majority of sounds. Only very few sounds produce statistically significant discrepancies in the two versions, which may be attributed to effects of the "meaning" of the sounds on the assessment of loudness.

1. Introduction

It is a well-recognized problem in psychoacoustics that *non-auditory* factors may enter into the evaluation of noises, or of the sound-quality of a test object [1, 2]. These influences may be due to attitudes towards the source, effects of familiarity, preferences, and user expectations about prototypical products. Such non-auditory influences on psychoacoustical judgments are sometimes summarized as effects of the *meaning* of the sound.

Recently Fastl [3] has proposed a signal-processing algorithm 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), and has recommended this algorithm to study the effects of meaning on the evaluation of sounds. 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, 4].

Initial psychoacoustic applications of the method to the continuous scaling of time-varying loudness [5], and to a selection of everyday sounds [6] proved to be encouraging; however, a large-scale investigation exploiting the potential of the method for conducting comparative listening tests is still lacking. Such a study should (1) use a wide variety of environmental and product sounds, (2) cover a large range of sound pressure levels, (3) include some measure of the success of the neutralization procedure (e.g. by determining the number of correct identifications), and (4) attempt to collect independent data on the meaning of the sounds (e.g. using the semantic differential technique) to facilitate the interpretation of discrepancies in judgments of original and processed sounds. The present report presents the first results from a larger study currently being conducted which fulfils most of these requirements.

2. Method

2.1. Participants

A total of 80 students at Aalborg University participated in the experiments. 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, they were randomly assigned to one of four conditions specified in Table 1, so that groups of 20 subjects having roughly equal gender composition were formed.

Table 1: *Experimental design: Tasks and stimuli.*

	Stimuli	
	original	neutralized
Loudness scaling	N=20 11 male / 9 fem.	N=20 11 male / 9 fem.
Annoyance scaling	N=20 10 male / 10 fem.	N=20 9 male / 11 fem.

From: ICA 2004, Kyoto, Intern. Commision on Acoustics

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 3 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), about a third of the sounds may be classified as product sounds of electrical devices recorded in their typical use. These sounds varied in duration from 0.7 to 5 s, and had overall sound-pressure levels between 30 and 80 dB SPL. In addition, seven levels of pink noise of 5 s duration, ranging from 20 to 80 dB SPL (in 10-dB steps) were included to check for the comparability of the subject groups.

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 loudness scaling task, the category subdivision procedure (CS, see [7]) 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 soft" (1-10), "soft" (11-20), "medium" (21-30), "loud" (31-40) and "very loud," (41-50). The endpoints of the resulting 50-point scale were verbally anchored to denote "inaudible" (0) and "painful" (beyond 50). After a short practice run, each subject judged the 47 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 this report we will focus on the loudness scaling data. Results on annoyance, and on the semantic differential technique will be analysed in a later report.

3.1. Pink noise reference

In order to check whether the two groups of subjects were comparable with respect to their loudness scaling behavior, an identical set of seven (unprocessed) pink-noise reference signals was interspersed, both among the original, and 'neutralized' sounds. Figure 1 shows that these references were judged nearly identically by the two groups of 20 subjects each, and that their judgments covered the entire range of the scale. An analysis of variance of these data shows a main effect of SPL, but no effect of group membership, and - most importantly - no group by SPL interaction (at $\alpha = 0.05$) which might have indicated a different growth of loudness for the two sets of listeners.

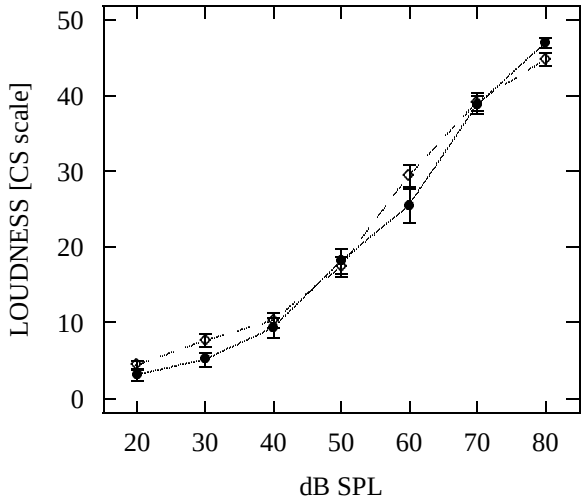

Figure 1: *Calibration curves: Loudness functions for identical samples of pink noise obtained from the two groups of subjects ($N = 20$ each) judging the original (filled circles) and neutralized (diamonds) sounds.*

3.2. Identifiability

The claim that the FTT procedure [3] obscures the *meaning* of a sound by reducing its identifiability may be evaluated by looking at the outcome of the identification task the subjects performed subsequent to the scaling experiment: For the present report, only the nouns assigned to the sounds were scored by comparing them with an a-priori list of acceptable answers. As is evident in Figure 2, FTT processing dramatically reduced the identifiability of the source from a median of 90% correct identifications with the original recordings to a median of 13% for the processed sounds. Note that the interquartile ranges associated with these numbers do not even overlap. The effect is by no means uniform across sounds, however, as discussed below.

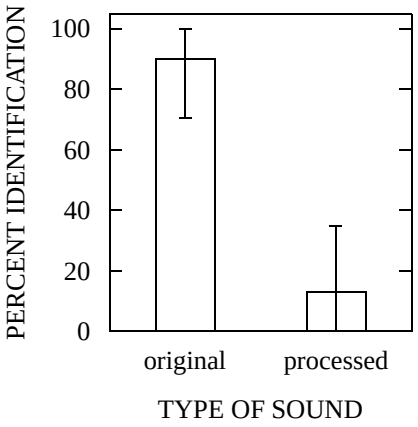

Figure 2: *Median correct identification of the 40 test sounds by the 20 listeners each participating in the two processing conditions. The error bars mark the upper and lower quartiles of the distributions.*

3.3. Loudness scaling

Figure 3 shows the outcome of the loudness scaling experiment, the sound samples being arranged in ascending order according to an instrumental loudness metric (mean statistical loudness; Brüel & Kjær PULSE sound-quality software type 7698). It is evident that judgments of the processed sounds (open diamonds) largely coincide with those of the original sounds (filled circles). A two-factor, mixed analysis of variance showed that in addition to the (higly significant, but trivial) main effect of the 40 sounds, there was no significant main effect of processing [$F(1, 38) = 0.215$; $p = 0.645$], confirming that there was no overall difference in loudness judgments between processed and unprocessed sounds. The fact that a significant (sound by processing) interaction [$F(39, 1428) = 4.55$; $p < 0.001$] is obtained, however, indicates that the *pattern* of mean loudness judgments depicted in Figure 3 significantly differs for original vs. FTT-processed sounds. Given the large number of sounds investigated, post-hoc tests as to where these differences occur have to be corrected for chance outcomes due to multiple testing: This was accomplished by computing Tukey's *honestly significant difference (HSD)*, i.e. a critical difference (in units of the loudness scale) that has to be exceeded to claim a significant effect. For the present loudness-scaling data, this difference turned out to be $HSD = 5.79$; that is - given the variability in the data, and the number of tests performed - a little more than what corresponds to half a verbal category. That difference was exceeded for three sounds only which are marked by arrows in Figure 3: The sound of an alarm clock, that of a buzzer, and the sound of a bicycle bell.

4. Discussion

The present study investigated a novel approach to obscuring the "meaning" of sounds [3, 4] in the context of loudness scaling. It employed a larger number of sounds, participants, and loudness levels, than previous applications of this approach. Furthermore, it used an identification task to empirically determine the effectiveness of the neutralization procedure.

The single most important novel finding in the present report may be to have shown that the "neutralization" procedure actually works: On the average, it drives down identifyability from 90% with the original recordings to a mere 13% for the processed sounds (see Figure 2). There is, however, a great deal of variance behind these statistics: Our sample contains sounds that are hard to identify in the first place (e.g. hairdryers, kitchen mixers), and others that remain identifyable even after "neutralization" (e.g. combustion motor sounds, the sound of knocking). What physical properties of the sounds determine their robustness towards "neutralization," will still have to be specified. Furthermore, even with incorrect identification, new 'meanings' of the sounds may emerge. Whether these are fuzzy, and idiosyncratic, or systematic across listeners, may be determined by further analysis of the identification, and semantic-differential data.

Furthermore, the loudness scaling data show, that the approach holds what it promises: The original and neutral sounds that are equivalent in instrumental loudness metrics via construction, are generally also *judged* to have equal loudness by human listeners. Note, however, that this conclusion is based on a very conservative use of inferential statistics: For a difference to become significant, it has to exceed half the width of a verbal category (5.79 scale units) on the 5-category scale used. The few significant discrepancies found occur with sounds for which "meaning" is plausible to have an effect: They are "signal" sounds (alarm, buzzer, bell) which typically require immediate action on the part of the listener. Why the effect points in the opposite direction for the sound of the bicycle bell (see Figure 3) is presently not accounted for.

The most convincing piece of evidence that we are actually dealing with effects of "meaning" comes from relating the identification data to the results of the loudness scaling experiment: When we select only those sounds for analysis for which "neutralization" worked optimally ($N = 12$), i.e. reduced identifyability from $\geq 80\%$ to $\leq 20\%$, we find a significant processing by sound interaction [$F(11, 418) = 5.05$; $p < 0.001$] as we did using the entire set of sounds. By contrast, using only those sounds ($N = 11$) for which neutralization reduced the identifiability by less than 30%, the interaction becomes insignificant [$F(10, 380) = 1.38$; $p = 0.20$], indicating that loudness differences between unprocessed and processed sounds only occur, when identifiability is greatly reduced.

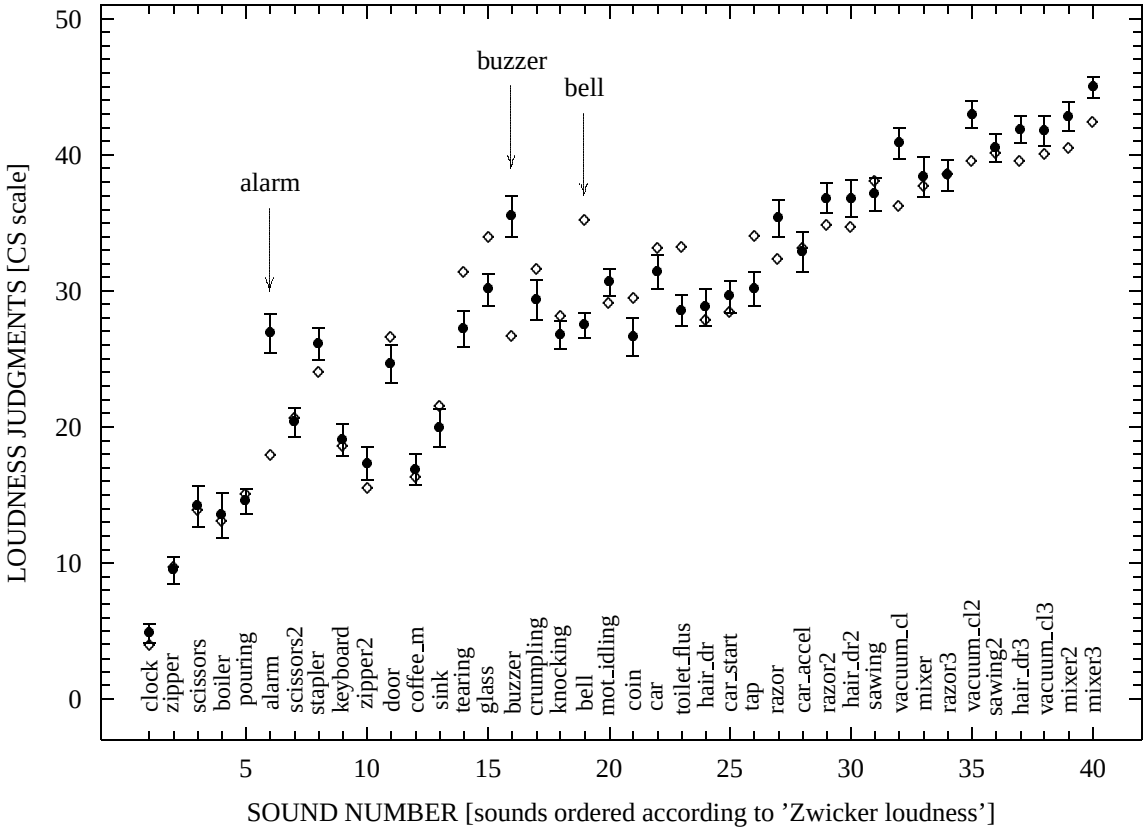

Figure 3: *Loudness judgments of the 40 test sounds (labelled along the abscissa) in their original (filled circles) and neutralized (diamonds) version. Variability is indicated for the original condition only by plotting the data points plus/minus one standard error of the mean; in the neutralized condition it is of comparable magnitude.*

Given the relative robustness of loudness judgments with respect to manipulations of identifiability, we will next consider what happens when annoyance, rather than loudness is being judged.

5. Acknowledgments

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 further like to thank Dipl.-Ing. Markus Fruhmann for processing the sounds for neutralization.

6. References

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

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

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

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

[5] J. Hellbrück, H. Fastl and B. Keller, "Effects of meaning of sound on loudness judgments," Proceedings Forum Acusticum 2002, Sevilla, Spain (CD-ROM).

[6] A. Zeitler, H. Fastl and J. Hellbrück, "Einfluss der Bedeutung auf die Lautstärkebeurteilung von Umweltgeräuschen," Proceedings DAGA 2002, Aachen, Germany (CD-ROM).

[7] 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 (317-336). Oldenburg, BIS, 1996.