Recognition of Environmental Sounds Embedded in Congruent and Incongruent Auditory Scenes

Jan Żera
Institute of Radioelectronics and Multimedia Technology,
Faculty of Electronics and Information Technology, Warsaw University of Technology,
Nowowiejska 15/19, 00-665 Warszawa, Poland

Tomira Rogala, Teresa Rościszewska, Joanna Szczepańska-Antosik, Andrzej Miśkiewicz, Jacek Majer
Department of Sound Engineering, The Fryderyk Chopin University of Music,
Okólnik 2, 00-368 Warszawa, Poland

Summary

Sound recognition accuracy and sound recognition time were measured in two experiments conducted with the use of recordings of environmental sounds. In Experiment 1 the sounds were mixed in natural auditory scenes and in Experiment 2 they were presented in quiet and in the background of a noise masker. The objective of the study was to examine whether musicians recognize the sources of environmental sounds more readily than non-musicians and to determine whether the speed of sound recognition is influenced by the contextual congruency of the sound and the background against which the sound is heard. The results show that musicians and non-musicians do not differ appreciably in sound recognition time and that the contextual congruency of the background scene has no effect on the speed of sound recognition. The finding of no difference between musicians and non-musicians in the speed of sound recognition is in contrast to what was expected from reports of enhanced auditory abilities of musicians in various non-musical listening tasks. The lack, in this study, of evidence for the "musician's hearing enhancement effect" is explained in reference to the typological categories of listening processes known as the modes of listening, distinguished in ecologically oriented auditory research.

PACS no. 43.66.Dc, 43.66Lj

1. Introduction

This paper reports a study conducted to compare the auditory abilities of musicians and non-musicians in the recognition of environmental sounds embedded in various acoustic background scenes. The specific purpose of the study was: (1) to determine whether musicians, owing to their refined hearing abilities, recognize the sources of environmental sounds more readily and more accurately than non-musicians, (2) to examine whether the speed of sound recognition depends on the contextual congruency of the background against which the target sound is heard.

The hypothesis assuming that musicians might possess more acute ability of sound recognition than non-musicians was inferred from a body of reports which indicate that musical training and practice develop not only strictly musical hearing, but also enhance a variety of non-musical auditory capabilities. The findings of what is called the "musicians' hearing enhancement effect" were obtained both in behavioral experiments [*e.g.*, 1, 2, 3] and in functional brain imaging studies [*e.g.*, 4, 5, 6].

In real life, individual sounds are usually heard in the background of other sounds. Published reports have shown that the listener's sound recognition ability may be facilitated or inhibited, depending on whether the sounds are presented in a cognitively congruent or incongruent context [7, 8, 9]. An auditory context, such as an acoustic background scene or a sequence of accompanying sounds, is congruent with the target sound when it is consistent with a real acoustic context in which the sound is encountered the environment. Although incongruent auditory scenes, created in laboratory experiments, are unrealistic, they have been used to explore various aspects of the process of sound recognition.

This study comprised two experiments conducted on separate groups of musicians and non-musicians.

In Exp. 1, environmental target sounds were mixed in the recordings of natural auditory scenes. In Exp. 2, the target sounds were played back in quiet and in the presence of a multitalker noise masker. In both experiments the listeners were instructed to give the responses immediately after recognizing the sound source.

2. Method

2.1. Target sounds and background scenes

The target sounds were recordings of 16 natural sounds. Recordings were made with a Neumann KU 100 dummy head, in two-channel stereo format. The sounds and their durations are listed in Table 1. The sounds were exemplars of three acoustic categories similar to those distinguished by Gygi *et al.* [10] in a typology of environmental sounds: harmonic sounds (sounds 1–6), impulsive sounds (7–11), non-harmonic continuous sounds (12–16). The full set of 16 sounds listed in Table 1 was used in Exp. 2. In Exp. 1 sounds 4, 9, 11, and 14 were omitted.

Table 1. Target sounds used in the study.

	Target sound	Duration (ms)
Harmonic Sounds		
1	bicycle bell	1296
2	bird calling	1442
3	car horn	637
4	laughter	1350
5	telephone ring	1376
6	whistle blow	738
Impulsive sounds		
7	computer typing	1049
8	coughing	724
9	door handle	906
10	footsteps	1445
11	glass breaking	673
Non-harmonic sounds		
12	car starting	1368
13	match lighting	809
14	toilet flushing	1457
15	water pouring	1245
16	zipper	702

In Exp. 1 the target sounds were mixed in an ongoing auditory scene played back at a loudness level of 65 phons. The auditory scenes were the following: (1) a busy street, (2) an indoor swimming pool, (3) a student cafeteria, (4) a large shopping hall. All scenes were recorded with a dummy head (Neumann KU 100) in two-channel stereo format. Each target sound was mixed at three signal levels in each scene. The lowest signal level was set such that the sound was just audible in the scene's background. The two other signal levels were by 3 and 6 dB higher than the lowest level. A series of sounds presented in the background of each scene comprised 36 stimuli (12 sounds × 3 signal levels). In Exp. 2 the target sounds were presented in two conditions: in quiet and in the background of a multitalker noise masker set at a loudness level of 65 phons. Each sound was presented at seven signal levels in each condition. The lowest level was close to the detection threshold and the other levels were by 6, 9, 12, 15, 18, and 24 dB higher. A series of sounds comprised 112 stimuli (7 sounds × 3 signal levels) in each condition.

2.2. Apparatus

The listening sessions were conducted individually with each subject, in a sound-attenuating booth. The set-up was built around two computers: a PC with an M-Audio Audiophile 2946 sound card and an iMac with an RME Fireface 400 interface. The PC monitor and the keyboard, placed in the booth, served as the listener's interface. The target sound files were read from the PC hard disk, mixed with the auditory scenes played back from the iMac's hard disk, and presented to the listener through a Beyerdynamic DT 990 headset. The listeners responded orally by naming the sounds they heard. Their responses, captured with a microphone in the booth, and the target sounds delivered to the headset were separately recorded on two audio tracks and stored on the iMac's hard disk.

2.3. Listening sessions

The listener was seated in front of a computer monitor in the booth and activated the presentation of a series of trials by pressing the space bar on the keyboard. The observation intervals within which the target sounds were presented in a series of trials were marked by a visual sign on the computer screen. Each observation interval lasted 10 s, regardless of the target sound's duration. The moment within the observation interval at which the target sound was switched on was chosen at random, with a reservation that the sound should terminate before the end of the interval. The listeners were instructed to speak out the response as soon as they recognized the sound and mark the response with a mouse on a list of target sounds

displayed on the computer screen. Entering the response on the screen activated a next test trial. When the listener was unable to recognize the sound source he/she gave a guessed answer.

2.4. Measurement of recognition time

The speed of sound recognition was determined by measuring the time interval between the onset of the target sound and the beginning of the listener's oral response. The duration of this interval is further called *recognition time* in this paper. The recognition time was read off-line from a screen display of two audio tracks simultaneously recorded during the session: a track with the target sounds and a track with a recording of the listener's oral responses.

2.5. Listeners

Experiments 1 and 2 were conducted on different groups of 10 musicians and 10 non-musicians. All listeners had pure-tone audiometric thresholds at 15 dB HL or less between 0.25 and 8 kHz. The musicians were students at the Fryderyk Chopin University of Music in Warsaw and the non-musicians were students from non-musical academic schools. None of the non-musicians had any experience in amateur musical activity. In each experiment and condition, a listener completed six repetitions of each series of the target sounds. The first series was a practice test and was omitted in the calculations of results.

3. Results and discussion

3.1. Experiment 1: Target sounds embedded in natural auditory scenes

The graphs in the left column in Fig. 1 show, for each target sound, the group mean recognition scores (mean percentage of correct recognitions) for musicians, plotted against the respective group scores of non-musicians. The graphs in the right column show a similar plot of the group recognition time. The data shown in Fig. 1 are example results obtained for four out of 12 target sounds. The different symbols in each graph indicate the auditory scene in which a given target sound was mixed.

It is apparent in the left group of panels in Fig. 1 that most data points lie above the diagonal dashed line on each graph which means that the recognition scores were generally higher for musicians than for non-musicians. A predominance of higher recognition scores in the musician group was also observed for the sounds not shown in Fig. 1, with an exception of the car honk, in the case of which the data points

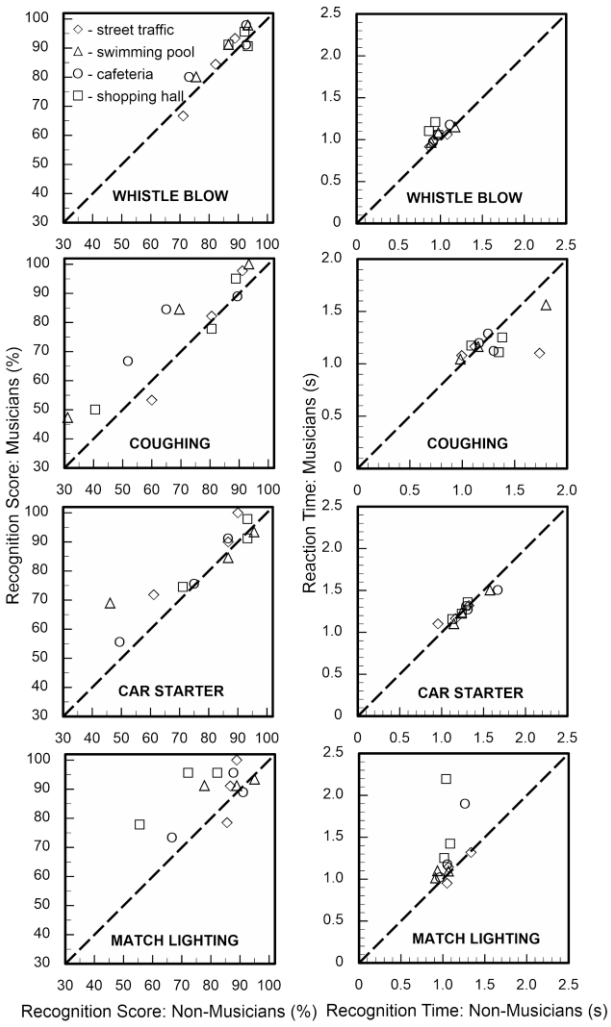

Figure 1. Group mean sound recognition scores and group median recognition time measured for musicians, plotted against the respective data for non-musicians. Individual panels present the data for one target sound, obtained at three signal levels, in four auditory scenes: street noise (diamonds), an indoor swimming pool (triangles), a student cafeteria (circles), and a large shopping hall (squares).

indicating the superiority of group of listeners over the other one were evenly distributed between the two groups. Overall, in 104 out of 144 data points in Exp. 1 the mean group sound recognition score was higher for musicians than for non-musicians and in five points the scores of both groups were equal. Although the group means suggest that there was an appreciable effect of musicianship on the ability of recognizing the sources of environmental sounds, the results of ANOVA have shown that the differences between the means of recognition scores calculated for musicians and non-musicians were not statistically significant in Exp. 1. The lack of statistical significance resulted from a small number of listeners and a relatively large dispersion

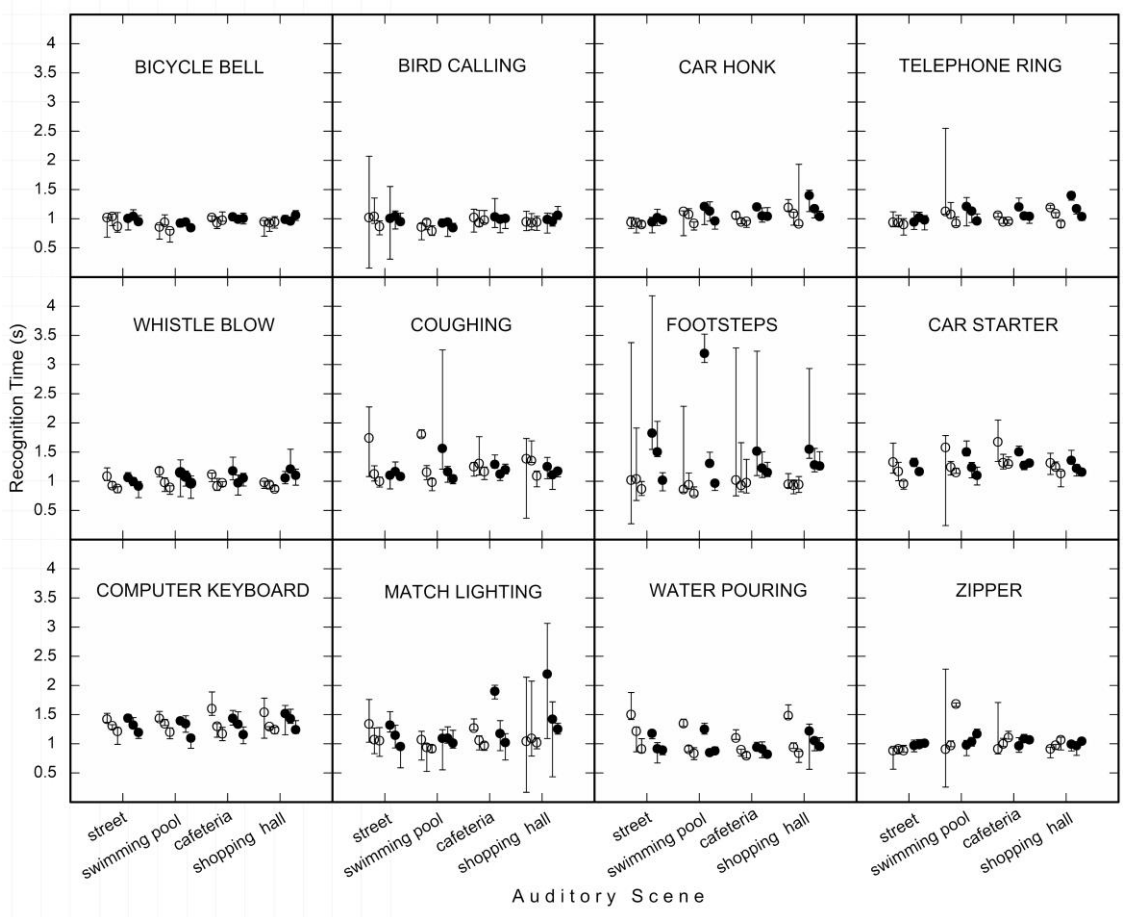

Figure 2. Sound recognition time for 12 target sounds mixed in four auditory scenes. Group medians calculated for non-musicians (open symbols) and musicians (closed symbols). The successive symbols grouped along the abscissa in each triad cluster show the data for three signal levels, in ascending order. The error bars indicate the interquartile ranges.

of data across the listeners in each group. In the case of most target sounds, musicians and non-musicians differed only very little, if at all, in the speed of sound recognition reflected by the sound recognition time. An exception is seen in the bottom right graph in Fig. 1: the recognition time determined for the sound of match lighting is much longer for musicians than for non-musicians. It also should be noted that the recognition scores obtained for the sound of match lighting, plotted in the adjacent panel on the left, were appreciably higher for musicians than for non-musicians. Improved sound recognition accuracy, obtained at the cost of prolonged recognition time, is symptom of speed-accuracy tradeoff, an effect widely known in psychophysics (see [11] for a review).

Figure 2 shows the sound recognition time for 12 target sounds in four auditory scenes. The data are group medians calculated for non-musicians (open symbols) and musicians (closed symbols). The successive symbols in each triad cluster, grouped along the abscissa, show the results for three signal levels, in ascending order. The error bars indicate the interquartile ranges.

The data in Fig. 2 show that sound recognition time decreases with increasing signal level and also depends, to some degree, on the typological acoustic category of the sound. Most of the group medians calculated for harmonic sounds (bicycle bell, bird calling, car honk, telephone ring, and whistle blow) are within a range of 0.8–1.1 s, with only a few slightly higher values at the lowest signal levels. For most impulsive and non-harmonic sounds the recognition time is somewhat longer and falls into a range of 0.8–2.0 s, with an exception of three results obtained for the sounds of match lighting and footsteps.

The finding that harmonic sounds were recognized slightly faster than non-harmonic sounds may be explained by the nature of auditory cues used in sound recognition. Most likely, the main cue for the recognition of harmonic sounds was their timbre associated with the character of the sound spectrum, whereas in the case of other sounds, the cues based

on the sound's temporal structure were presumably more important.

When a sound is recognized upon its temporal structure more time may be needed for extracting such a type of information from the stimulus than in the case of stationary spectral cues. In many cases an increase in recognition time may be therefore an effect of the sound's temporal structure and may not always indicate that the cognitive processing of sensory information was prolonged.

3.2. Experiment 2: Target sounds presented in quiet and against a noise masker

Figure 3 presents example psychometric functions for the recognition of environmental sounds in quiet (open symbols) and in the presence of a multitalker noise masker (closed symbols). The functions, determined separately for musicians (circles) and non-musicians (squares), show the group mean percentage of correct recognitions plotted against the sound exposure level of the target sound.

The most noteworthy finding apparent in Fig. 3 is that the psychometric functions are nearly identical for musicians and non-musicians which means that those groups of listeners did not differ in sound recognition accuracy. Similar, close convergence of sound recognition psychometric functions determined for musicians and non-musicians was also observed in the case of other target sounds, not shown in Fig. 3 and the functions had a similar slope across the target sounds.

Figure 4 shows, in a separate panel for each target sound, the group median sound recognition time at seven signal levels, represented by the sound exposure level on the abscissa. The data, measured for musicians (circles) and non-musicians (squares), are shown for target sounds played back in quiet (open symbols) and in the background of a continuous multitalker noise masker (squares).

It is readily apparent in Fig. 4 that, both in quiet and in masked conditions, the recognition time markedly decreased with increasing signal level. For most sounds the recognition time ranged from about 0.8–1.0 s at the highest signal level to about 2–3 s at the lowest level. In the case of two target sounds (bird calling and door handle) the group median recognition time amounted to 3.4 s at the lowest signal level, in the group of non-musicians.

An important observation apparent in Fig. 4 is that at high signal levels recognition time is about the same for all target sounds, conditions of presentation, and groups of listeners. The differences across sounds are only seen at low signal levels but they do not demonstrate any clear-cut relation of sound recognition time to the typological categories of environmental sounds distinguished in the literature [10]. The recognition time values plotted in Fig. 4, measured at high signal levels for individual sounds, agree fairly well with the data shown in Fig. 2, obtained for the same target sounds embedded in natural auditory scenes. At low signal levels such a convergence of results is not observed as the recognition time values are longer in Fig. 4 than those shown in Fig. 2. The reason for such a divergence of recognition time observed for the same sounds in different experiments is unclear. It should be, however, noted here that during the preparation of Exp. 1 the lowest signal level was only roughly estimated for each target sound and auditory scene. Knowing that recognition time decreases with increasing signal level and increasing loudness one may presume that the sounds presented at low signal levels in the background of natural auditory scenes in Exp. 1 were somewhat louder than low-level sounds in Exp. 2 and therefore yielded shorter recognition time.

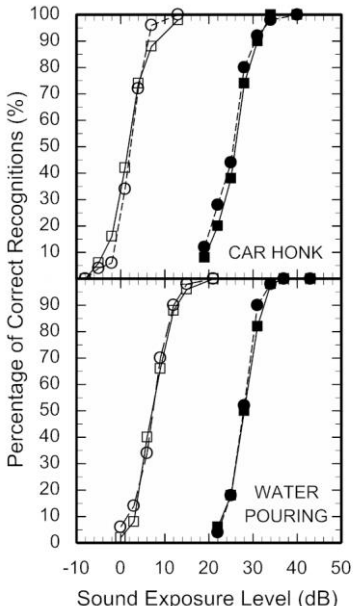

Figure 3. Psychometric functions for sound recognition determined for target sounds in quiet (open symbols) and in the background of a multitalker noise masker (closed symbols). The functions are shown separately for musicians (circles) and non-musicians (squares).

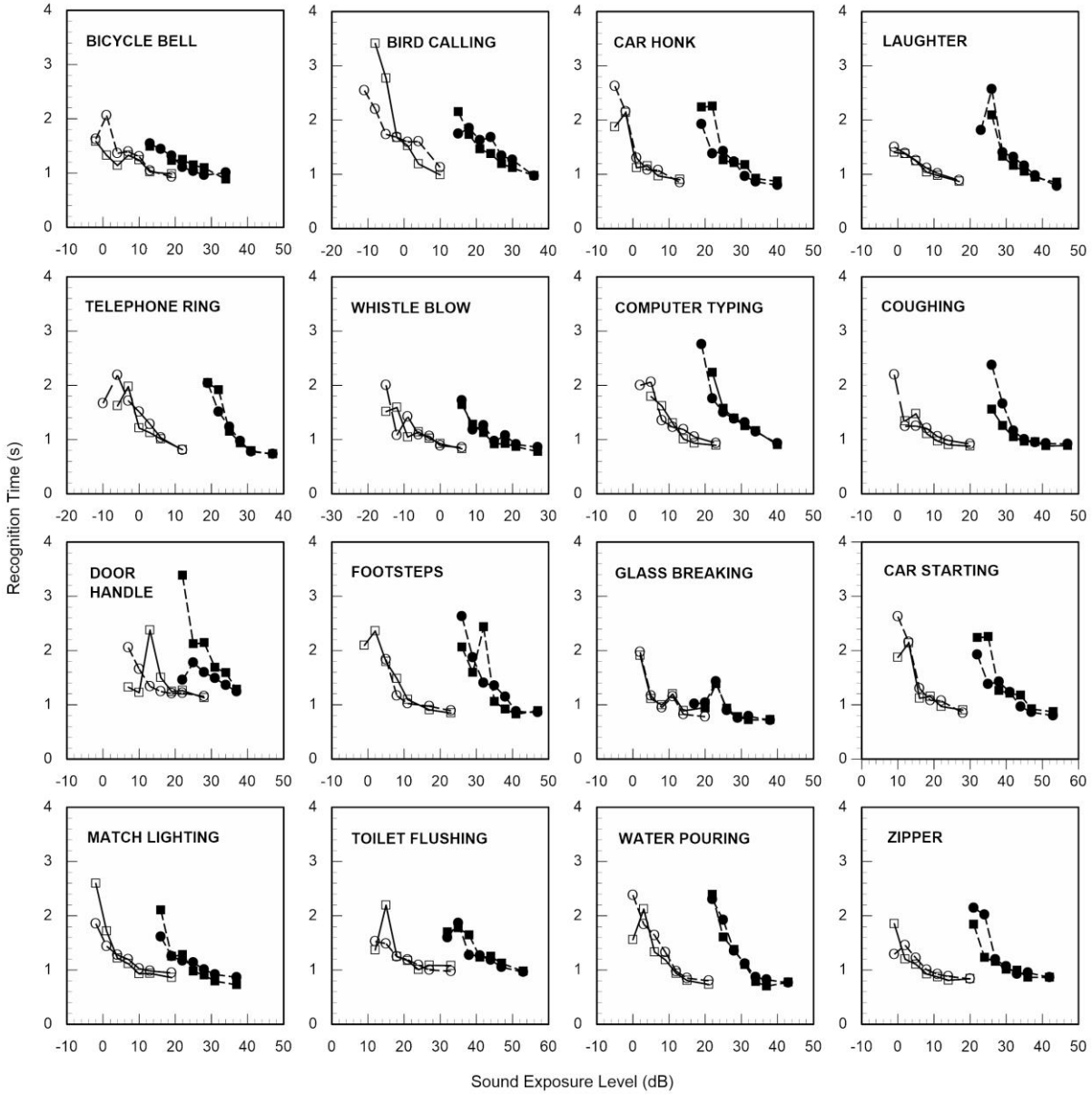

Figure 4. Group median recognition time measured in Exp. 2 for individual target sounds in quiet (open symbols) and in the background of multitalker masking noise (closed symbols). The data are shown separately for musicians (circles) and non-musicians (squares). The abscissa is the sound exposure level of target sounds.

The main research problem, put forward in Exp. 2, was whether musicians recognize the sources of environmental sounds more readily than non-musicians. The data shown for 16 target sounds in Fig. 4 do not provide any evidence that musicians and non-musicians might differ in the speed of sound recognition. The great majority of data obtained in Exp. 2 for musicians and non-musicians is in very close agreement and an appreciable difference between the results of those groups is apparent only in a few data points in Fig. 4. This finding is in agreement with the results of Exp. 1 in which no difference was observed between musicians and non-musicians in the speed of recognition of environmental sounds in natural auditory scenes.

4. General discussion

Discussion of the results should, above all, address the two main research questions of the study and resolve: (1) whether there is an appreciable difference between musicians and non-musicians in the accuracy and the speed of recognizing the sources of environmental sounds, (2) whether the sound recognition speed is influenced by the contextual congruency of the target sound and the acoustic background against which the sound is heard.

The present experiments did not reveal any appreciable superiority of musicians over non-musicians in the ability of recognizing the sources of environmental sounds. Although musicians obtained, by and large, higher sound recognition scores than non-musicians, the difference in sound recognition accuracy was not pronounced enough strongly between those two groups of listeners to yield statistically significant results. Furthermore, the measurements of recognition time have shown that musicians and non-musicians do not differ in the speed of sound recognition

The finding of no appreciable effect of musicianship on the ability of recognizing the sources of environmental sounds is in agreement with the results of our earlier measurements of recognition-detection threshold gaps for environmental sounds [12]. Those results have shown that the recognition-detection threshold gap, defined as the minimum sound level above detection threshold, at which the listener is able to recognize a given sound source, is about the same for musicians and non-musicians. This finding indicates that, in particularly perceptually demanding conditions, when the sounds are presented at near-threshold levels and not all their acoustic signatures are clearly audible, musicians do not recognize the sources of environmental sounds more accurately than non-musicians.

The finding of no superiority of musicians over non-musicians in the ability of recognizing the sources of environmental sounds is in contrast with what might be inferred from the reports of the phenomenon known as the "musicians' hearing enhancement." One possible explanation of the lack of evidence for such a phenomenon in the current study is that the musicians' hearing enhancement effect is observed only in certain modes of listening. The term "mode of listening" refers to the specific listening strategy used by a person in an auditory task [13, 14, 15, 16]. Usually, three basic modes of listening are distinguished in their typology: causal listening, semantic listening, and reduced listening [15]. *Causal listening*, also termed *everyday listening* [14, 15], is focused on auditory orientation in the listener's surrounding environment. The focus of *semantic listening* is to extract the information conveyed by the sounds by means of a certain code or language. The objective of *reduced listening*, also termed *musical listening* [14, 15], is to perceive the inherent sonic attributes of sound with no connotations to any sound sources or events that might produce the sounds.

The findings of the musician's hearing enhancement were obtained in experiments in which the subjects listened to the test sounds either in the semantic mode or in the reduced mode, whereas recognition of environmental sounds, explored in the present study, belongs to the category of causal listening. The current study, as well as our earlier experiments [12] have shown that the listening skills and abilities associated with the causal mode of listening are only very little, if at all, improved by musicianship.

The present study did not provide any evidence for the influence of the contextual congruency of the target sound and the background on the acuity of sound recognition. It should be, however, noted that the reports of such an effect were based on a comparison of the percentage of correct responses obtained for the same target sounds mixed with different background scenes, at the same signal-to-noise ratio [8, 9]. A shortcoming of such an inference is that the recognition scores compared at the same S/N for different target sounds and auditory scenes also depend on the amount of masking produced by the background scene. At a given S/N ratio, masking may vary, depending on the spectral and temporal characteristics or the target sound and the background. To verify the findings of the effect of contextual congruency on sound recognition we compared the recognition time measured for the same sounds in different scenes, instead of comparing the percentage of correct recognitions obtained at a given S/N ratio. The measurement of recognition time was introduced by Gordon [17] for the assessment of contextual effects in visual perception.

One may hypothesize that the lack in the current study of evidence for the influence of contextual congruency on the sound recognition time might be a result of the procedure of sound presentation. The target sounds were presented in Exp. 1 against a continuous recording of an auditory scene, so that the listeners were immersed in the acoustic content of the scene during the session. Such a perceptual immersion in the background sound could possibly facilitate the recognition of target sounds. It should be noted here that the reported findings of the contextual congruency effect in sound recognition were obtained for target sounds mixed with very short excerpts of auditory scenes and the scenes were changed from trial to trial in the experiment [8, 9].

A noteworthy observation, apparent both in Exp. 1 and Exp. 2, is a decrease in sound recognition time with increasing signal level of the sound. It has long

been known that human response time decreases in stimulus detection tasks as a power function of stimulus intensity. This relationship has been known as Piéron's law [18]. Over the past century Piéron's law had been studied in relation to various sensory modalities, including the auditory modality [19]. More recent studies of visual perception have shown that Piéron's law also holds in multiple choice tasks [20, 21]. The present finding of the influence of signal level on the sound recognition time suggests that Piéron's law possibly applies to stimulus recognition in the auditory modality.

5. Conclusions

The main findings of the current study may be summarized as follows.
(1) Musicians and non-musicians possess similar ability of recognizing environmental sounds.
(2) The finding of no appreciable superiority of musicians over non-musicians in sound recognition is in contrast with the reports of the musicians' hearing enhancement effect. The lack of evidence for such an effect in the current study is explained in terms of the listening mode representing the listening strategy of sound recognition. Recognition of environmental sounds is based upon the causal listening mode whereas the enhanced hearing abilities of musicians were observed in experiments concerned with the reduced and with the semantic modes of listening.
(3) Piéron's law, concerned with reaction time in stimulus detection, also seems to be applicable to sound recognition multiple choice tasks.

Acknowledgements

This work was supported by the National Science Centre Poland, Grant UMO-2013/11/B/HS6/01252, *Recognition of environmental sounds by musicians and non-musicians.* The first author's participation in the Euronoise 2018 congress was supported by a grant 504G/1034/0416 from the Warsaw University of Technology**.**

References

[1] A. J. Oxenham, B. J. Fligor, C. R. Mason, G. Kidd Jr: Informational masking and musical training. J Acoust Soc Am 114 (2003) 1543–1549.

[2] A. Parbery-Clark, E. Skoe, C. Lam, N. Kraus: Musician enhancement for speech-in-noise. *Ear and Hear* 30 (2009) 653–661.

[3] D. L. Strait, N. Kraus, A. Parbery-Clark, R. Ashley: Musical experience shapes top-down auditory mechanisms: Evidence from masking and auditory attention performance. Hear Res 261 (2010) 22–29.

[4] G. Musacchia, M. Sams, E. Skoe, N. Kraus: Musicians have enhanced subcortical auditory and audiovisual processing of speech and music. Proc Nat Acad Sci 104 (2007) 15894–15898.

[5] C. Pantev, B. Ross, T. Fujioka, L. J. Trainor, M. Schutte, M. Schulz: Music and learning-induced cortical plasticity. Ann NY Acad Sci 999 (2007), 438–450.

[6] S. C. Herholz, B. Boh, C. Pantev: Musical training modulates encoding of higher-order regularities in the auditory cortex. Eur J Neurosci 34 (2011) 524–529.

[7] J. A. Ballas, T. Mullins: Effects of context on the identification of everyday sounds. Hum Perform 4 (1991) 199–219.

[8] R. Leech, B. Gygi, J. Aydelott, F. Dick: Informational factors in identifying environmental sounds in natural auditory scenes. J Acoust Soc Am 126 (2009) 3147–3155.

[9] B. Gygi, V. Shafiro: The incongruency advantage for environmental sounds presented in natural auditory scenes. J Exp Psychol Hum Percept Perf 37 (2011) 551–565.

[10] B. Gygi, G. R. Kidd, C. S. Watson: Similarity and categorization of environmental sounds. *Percept Psych* 69 (2007) 839–855.

[11] R. P. Heitz: The speed-accuracy tradeoff: History, physiology, methodology, and behavior. Front Neurosci 8 (2014), article 150.

[12] T. Rościszewska, A. Miśkiewicz, J. Żera, J. Majer, B. Okoń-Makowska: Detection and recognition thresholds of environmental sounds. In: M. Meissner (Ed.) Advances in Acoustics. Warszawa, 2016.

[13] W. W. Gaver: What in the world do we hear? An ecological approach to auditory event perception, Ecol Psychol 5 (1993) 1–29.

[14] W. W. Gaver: How do we hear in the world? Explorations in ecological acoustics. Ecol Psychol 5 (1993) 285–313.

[15] M. Chion: Audio-Vision: Sound on Screen. Columbia University Press, New York (1994).

[16] A. Preis, A. Klawiter: The audition of natural sounds – its levels and relevant experiments. Proc. Forum Acusticum, Kraków (2005) 1595–1599.

[17] R. D. Gordon: Attentional allocation during the perception of scenes. J Exp Psychol Hum Percept Perform 30 (2004), 760–777.

[18] H. Piéron: Recherches sur les lois de variation des temps de latence sensorielle en fonction des intensités excitatrices. L'Année Psycholog 20 (1914) 17–96.

[19] R. Chocholle: Variation des temps de reaction auditifs en function de l'intensite à diverses fréquences. Année Psychol 41 (1940) 65–124.

[20] J. Palmer, A. C. Huk, M. N. Shadlen: The effect of stimulus strength on the speed and accuracy of a perceptual decision. J Vis 5 (2005) 376–404.

[21] T. Stafford, L. Ingram, K. N. Gurney: Piéron's law holds during stroop conflict: insights into the architecture of decision making. Cogn Sci 35 (2011) 1553–1566.