



Effects of Noise on Speech Perception and Spoken Word Comprehension

Jovan Eranović¹, Daniel Pape¹, Magda Stroinski¹, Elisabet Service¹, Marijana Matkovski²

¹McMaster University, Canada

²University of Novi Sad, Serbia

{eranovij, paped, stroinsk, eservic}@mcmaster.ca, marijanamatkovski@uns.ac.rs

Abstract

The aim of the study was to find out which of the three categories of noise acting as maskers (*energetic*: masking portions of the target speech with its energy; *informational*: both target and masker compete for the listener's attention; *degraded*: reverberated or filtered speech) is most detrimental to speech perception and spoken word comprehension. To that end, participants completed three tasks with and without added noise – listening span, listening comprehension, and shadowing – where shadowing is considered primarily a task relying on speech perception, with the other two tasks considered to rely on word comprehension and semantic inference. The study found informational masking to be most detrimental to speech perception, while energetic masking and sound degradation were most detrimental to spoken word comprehension. The results also imply that masking categories must be used with caution, since not all maskers belonging to one category had the same effect on performance.

Index Terms: speech in noise, energetic masking, informational masking, speech perception, word comprehension

1. Introduction

Speech perception is a process of identification of auditorily presented phonemes, which includes interaction of these phonemes with each other as well as their matching with phonemic representations in our long-term memory, and finally their interpretation [1]. Spoken word comprehension is the listener's ability to arrive at the intended meaning [2]. While speech perception is an integral part of spoken word comprehension, the process itself can be carried out without paying much attention to overall meaning. In a typical conversational environment, our ability to clearly and accurately perceive speech relies on the ability of our auditory systems to distinguish incoming speech from the background noise [3]. The term *auditory masking* describes deficits in sound perception, and it is defined as “the process by which the threshold of hearing one sound is raised by the presence of another” [4, p.110]. There are three commonly discussed types of noise causing mismatch between what listeners expect to hear and the actual acoustic signal they receive, and resulting in heavier cognitive loads, slower processing, and impaired messages.

1.1. Common Types of Auditory Maskers

Noise can act as an energetic masker covering portions of the frequency spectrum, or an informational masker confusing the listener in the decision as to which sound source to attend to. Energetic masking refers to a process in which the intensity and frequency of background noise render parts of the target speech

signal inaudible [5]. Informational masking occurs in listening situations in which the competing non-target speech is perfectly audible – thus acting as a distractor, making the listener unable to separate the elements of one from another [5]. Erroneously considered to be a background noise, masking energy in reverberation does not come from an outside source, but rather originates in the target speech itself, reflecting off nearby surfaces [6]. Reverberated speech is the original speech signal combined with its time-delayed reflections that typically result in a smeared signal [7].

1.2. Tasks

Using three tasks – listening span, listening comprehension, and shadowing – the study aims to find out to what extent different types of background noise affect speech perception and spoken word comprehension. The listening span task is used for measuring the maximum amount of information a person's short-term memory can store [8]. In a typical experimental setting, subjects are presented with a sequence of auditory stimuli, and asked to recall specific items either in free or serial order. The maximum number of items correctly recalled determines one's listening span. Listening comprehension is a process during which listeners formulate meaning of the input speech based on their linguistic competence and contextual cues. The listening comprehension task assesses one's ability to understand material presented in auditory mode, make inferences, and draw logical conclusions. Shadowing is an activity during which subjects simultaneously listen and repeat the model's speech, trying to reproduce the phonological representations from the perceived auditory input [9]. It is used to study selective attention and memory in humans [10], but also an effective tool for improving listening skills [11].

1.3. Research Questions

The study investigates the effects of noise on speech perception and spoken word comprehension. Specifically, the aim of the study is to answer the following two research questions: (1) Are speech perception and spoken word comprehension equally affected by noise maskers? (2) What type of commonly occurring noise maskers has the most detrimental effect on speech perception and word comprehension?

2. Method

2.1. Participants

Participants were recruited through SONA linguistics research participation system at McMaster University. Fifty undergraduate students (academic years 1-3; 33 females and 17 males) participated in the study for class credit. All participants were native speakers of English and reported normal hearing at the time of testing. None of the participants spoke the Mandarin

or Greek language (see *single babble* masker and *multi babble* masker definitions below).

2.2. Stimuli and Procedure

In each experiment participants listened to 7 spoken audio clips – one containing no background noise (*clean* condition), and the remaining 6 containing either informational or energetic maskers– for a total of 21 different speech stimuli across the three experiments. All the clips were edits culled from an exercises audio CD that accompanies *Cambridge Vocabulary for IELTS Advanced Band 6.5+* [12]. Noise maskers added to the clips were taken from the BBC Sound Effects Library [13]. All the maskers were added at the signal-to-noise ratio of -5 dB, based on the results of previous studies which found that this particular ratio kept “the average intelligibility in the 45%-65% range” [14]. Considered to be moderate, this ratio was preferred to “ensure that listeners would not perform at ceiling in an ‘easy’ listening condition or at the floor at the more difficult SNR” [15]. The energetic maskers used were *construction* noise (const. - drills and jackhammers), *single babble* masker (s.b.e. – news broadcast in Mandarin, with no embedded music), and *multi babble* masker (m.b.e. – 4-voice bar chatter in Greek). Mandarin and Greek were used as energetic maskers for none of the participants were familiar with these languages; otherwise they would have acted as informational maskers. The informational masker (s.b.i.) was a news broadcast in English. The clips that featured degraded sound were made by narrowing down the speech frequency bandwidth to 350-3400 Hz range (in case of the *phone* condition), or creating a reverb effect (in case of the *reverb* condition). The reverberation time was 1s (pre-delay 47 ms) – values typically found in classrooms with unfavorable acoustics and larger conference rooms [16]. The mixing and effects were done using Pro Tools 2020.5. The audio files used were sampled at 44.1 kHz with 16 bits and normalized for loudness. In order to achieve randomness, the stimuli were counterbalanced across conditions – after every ten participants noise effects were shuffled across the speech tracks so that the next group would hear different speech tracks accompanied by a different noise source. The experiments were conducted online via the Zoom software. The stimuli in the listening span and listening comprehension tasks were played from the researcher’s computer via the *Share Audio* function in Zoom. Participants were instructed to turn off their cameras as well as any background applications that could negatively influence memory and Internet bandwidth, and keep the Zoom audio setting set to *high fidelity music* mode. This ensured that no decrease in audio bitrates was expected to occur, which might have affected the results. The researcher’s computer was connected to the Internet via ethernet. Importantly, no participant reported any connectivity issues during the experiments, nor Zoom itself issued any warnings. Participants listened to the stimuli through their own wired headphones set at a comfortable (listener-adjusted) level. In the third task, participants recorded their own voices using their mobile phones. They were instructed to keep the phones at the distance of approximately 60cm, in order to avoid signal distortion. Responses were not timed.

2.3. Experiments

Participants performed all three experiments in one session, with a mean time of around 45-50 minutes per participant – with short breaks between the tasks. The experiments always proceeded in the following order: listening span task, listening comprehension task, shadowing. Participants received

instructions prior to each experiment, and they also completed two or three practice trials in order to become familiar with the task, and adjust the volume of their headphones. Practice trials used no background noise. Experimental trials were shuffled across participants, so that the order in which they were exposed to different types of noise was counterbalanced. In all three experiments, one clip contained no background noise (*clean* condition), while the rest contained six different types of noise maskers described above. All three experiments were a single-task design, with no other data collected.

2.3.1. Tasks

Listening Span: Participants listened to seven five-sentence audio clips (31-42s long; mean = 36.5s; sd = 7.77). They were instructed to remember the last word of each sentence in order of presentation, thus having to remember five target words per condition.

Listening Comprehension: Participants listened to seven audio clips (62-92s long; mean = 77s; sd = 21.21), after which they were asked to answer four multiple-choice questions related to the material they heard.

Shadowing: *Shadowing* is defined as “a paced auditory tracking task which involves the immediate vocalization of auditorily presented stimuli, *i.e.*, word-for-word repetition, *in the same language*, parrot-style, of a message presented through headphones” [17]. Participants shadowed seven audio clips (59-72s long; mean = 65s; sd = 8.48). In order to avoid any technical difficulties typical for online environments (such as echoes, delays, sound errors due to unstable internet connection, etc.), participants were emailed the shadowing stimuli immediately prior to the experiment, and they played the stimuli from their own computers. They were instructed to delete the stimuli after the experiment.

3. Results

3.1.1. Listening Span Task

The mean, median, error percentage, and standard deviation were calculated for each condition, as given in Table 1.

Table 1: Mean, median, mean percentage of errors and standard deviation in the listening span task

condition	mean	median	%	sd
clean	3.62	4.0	72.4	1.40
const.	2.76	3.0	55.2	1.45
m.b.e.	3.60	4.0	72.0	1.41
phone	3.88	4.0	77.6	1.14
reverb	3.88	4.0	77.6	1.42
s.b.e.	3.64	4.0	72.8	1.41
s.b.i.	3.24	3.0	64.8	1.41

A Shapiro-Wilk test for normality found that none of the experimental condition showed normal distribution, so the Wilcoxon non-parametric paired tests for significance were conducted for each noise condition compared to the clean signal baseline. The tests found no statistically significant difference in medians except for construction noise, meaning that none of the other types of background noise affected the task results. Interestingly, a reverse statistically significant difference between the clean and construction conditions was found, as shown in Table 2.

Table 2: Results of the Wilcoxon test in the listening span task (incl. Bonferroni correction)

condition	p-value	z-score	p-adj
clean -const.	0.0007351	-3.376159	0.0044106**
clean - m.b.e.	0.8087	-0.242152	1.0000000
clean - phone	0.4528	-0.750698	1.0000000
clean - reverb	0.2467	-1.158385	1.0000000
clean - s.b.e.	0.7335	-0.340459	1.0000000
clean - s.b.i.	0.09875	-1.650924	0.5925000

Finally, the between-group comparison found statistically significant differences in error distribution between degradation vs energetic and degradation vs informational.

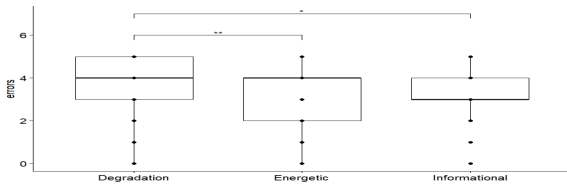


Figure 1: Statistically significant differences established by the pairwise Wilcoxon signed-rank test

3.1.2. Listening Comprehension Task

The mean, median, error percentage, and standard deviation were calculated for each condition, as given in Table 3.

Table 3: Mean, median, mean percentage of errors and standard deviation in the listening comprehension task

condition	mean	median	%	sd
clean	1.34	1.0	33.5	0.75
const.	2.62	3.0	65.5	0.90
m.b.e.	1.56	1.5	39.0	0.93
phone	2.72	3.0	68.0	0.90
reverb	1.80	1.0	45.0	0.90
s.b.e.	2.16	2.0	54.0	0.84
s.b.i.	1.80	2.0	45.0	1.03

The Shapiro-Wilk test found that none of the experimental conditions had normal distribution, so again Wilcoxon tests for significance were conducted. Table 4 shows that when compared with the control (clean signal) statistically significant differences in medians were found in all noise conditions, except for the multi babble energetic, which was the only condition that did not affect the performance.

Table 4: Results of the Wilcoxon test in the listening comprehension task (incl. Bonferroni correction)

condition	p-value	z-score	p-adj
clean - const.	0.0000002061	-5.193769	0.0000012***
clean - m.b.e.	0.1635	-1.393406	0.98100
clean - phone	0.0000000645	-5.405777	0.0000003***
clean - reverb	0.005418	-2.781099	0.032508
clean - s.b.e.	0.0000006952	-4.495166	0.000041***
clean - s.b.i.	0.002622	-3.008873	0.015732

Finally, the between-group comparison found statistically significant differences in error distribution between degradation and informational.

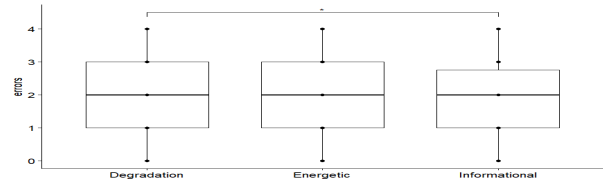


Figure 2: Statistically significant differences established by the pairwise Wilcoxon signed-rank test

3.1.3. Shadowing

Since the number of possible errors in this experiment varied across the stimuli (mean = 157.14, sd = 12.18), the raw data was first scaled to 0-100% range, after which the statistical analysis was performed. The mean, median, mean error percentage, and standard deviation were calculated for each condition, as given in Table 5.

Table 5: Mean, median, mean percentage of errors and standard deviation in the shadowing task

condition	mean	median	%	sd
clean	3.84	3.0	3.84	3.87
const.	7.85	6.2	7.85	7.61
m.b.e.	7.48	5.7	7.48	6.51
phone	5.68	4.0	5.68	5.04
reverb	9.62	6.7	9.62	8.68
s.b.e.	5.14	3.7	5.14	5.32
s.b.i.	10.39	7.9	10.39	8.74

The Shapiro-Wilk test found that none of the experimental conditions had normal distribution, so the Wilcoxon non-parametric paired tests for significance were conducted to determine which types of background noise significantly affected the results of the shadowing task (Table 6).

Table 6: Results of the Wilcoxon matched-pairs signed rank test in the shadowing task (incl. Bonferroni correction)

condition	p-value	z-score	p-adj
clean -const.	0.0000853	-3.928999	0.0005118***
clean - m.b.e.	0.00003756	-4.121998	0.0002253***
clean - phone	0.005347	-2.785346	0.032082
clean - reverb	0.000000851	-4.923258	0.0000051***
clean - s.b.e.	0.04175	-2.035978	0.25050
clean - s.b.i.	0.000000027	-5.560321	0.00000016***

Finally, the between-group comparison found statistically significant differences in error distribution between energetic and informational maskers.

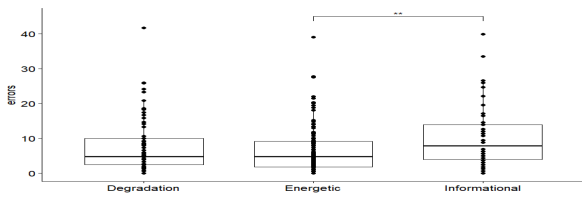


Figure 3: *Statistically significant differences established by the pairwise Wilcoxon signed-rank test*

4. Discussion

Overall, the results show that participants found the listening span task quite challenging (Table 1), with an average error rate of 71% across all conditions noticeably high. The obtained results correspond with the nature of a serial recall task, known to be quite demanding on the short-term memory [18]. No statistically significant difference in error rate was found across conditions, except for the construction masker, which, very surprisingly, yielded 17% fewer errors than the control. Since the speech material was counterbalanced and loudness-normalized we are confident that this result is not a direct result of our acoustic/speech stimuli design. Task replication is recommended to hopefully account for this occurrence, as well as for the unexpectedly poor scores in the clean condition. Results also show that in single babble energetic and multi babble energetic conditions participants made 73% and 72% errors respectively. While informational maskers are typically thought of as more detrimental to speech perception than their energetic counterparts, in this experiment informational masking yielded 6% fewer errors than energetic maskers. Both noise conditions from the degradation group – reverb and phone – were found to result in the greatest percentage of errors at 78%. For the reverb condition, this is consistent with Rogers and colleagues who found the performance of their participants considerably poorer in the reverb condition [6]. For the phone condition, such score was anticipated since it has been well established that “the reduced bandwidth of the telephone speech accounts for a significant amount of performance deterioration” [19, p. 189]. Overall, the findings suggest that when engaged in a relatively short memory single-task, such as listening span, with no embedded or simultaneous secondary tasks, short-term memory capacity is equally unaffected by both energetic and informational maskers, as well as degraded speech. Between-group analysis found significant statistical difference between signal degradation and energetic masking, as well as between signal degradation and informational masking (Figure 1).

The listening comprehension task tested participants’ ability to understand auditorily presented stimuli in adverse conditions, and make inferences about the ideas discussed in the recordings presented in adverse conditions. Detrimental effect of background noise to listening comprehension had been documented in previous studies [20], [21]. With that in mind, it has been predicted that the outcomes of the listening comprehension test would be affected by background noise maskers. The results show the average of 50%, error rate across all conditions (Table 4). Statistically significant differences were found in two energetic maskers as well as in the phone condition. The greatest number of errors was measured in the phone condition, which corresponds with the idea that quality of speech perception suffers dramatically when high frequencies are removed [22], [23]. This also implies that

relevant linguistic information is contained in the high-frequency band, which should be tested and expanded upon in future research. The two energetic maskers that most affected listener comprehension were construction noise and single babble, with error rates of 66% and 54% respectively. Surprisingly, the single babble informational masker (English language) yielded 9% fewer errors than the single babble energetic (foreign language), in contrast to what has previously been reported [24], [25]. Even though statistically significant, the reverb condition proved not to be too detrimental in this particular task, with only 12% more errors than the control. Overall, noise maskers do have negative effect on listening comprehension. Between-group analysis found significant statistical difference between signal degradation and informational masking (Figure 2).

In the shadowing task, statistically significant differences were found in all experimental conditions, except the phone and single babble energetic. The greatest number of errors (10%) was found in the single babble informational condition, while the single babble energetic masker was found to affect performance less (5%) than its informational counterpart. The two remaining energetic maskers – construction and multi babble energetic – resulted in 8% and 7% error rate respectively. A high error rate of 10% found in reverb was only 1% less detrimental than the single babble informational masker. Finally, the phone condition yielded an error rate of 6%, resulting in 2% more errors comparing with the clean condition. The obtained results correspond with previous findings showing that noise maskers significantly affect speech perception in noise [6], [26], [27]. Between-group analysis found significant statistical difference between energetic and informational masking (Figure 3). To the authors’ knowledge, no study so far tested shadowing in noise in a single-task design.

5. Conclusions

Both speech perception and spoken-word recognition are processes relying primarily on working memory, which gets particularly taxed in adverse listening environments. The results show that the two processes are not equally affected by noise maskers, indicating that informational masking is most detrimental to speech perception, while energetic masking and sound degradation are most detrimental to spoken word comprehension. In addition, the study concludes that categories of energetic masking, informational masking, and degraded speech must be used with caution, since not all maskers belonging to one category have the same effect on performance. The study cannot offer any reliable predictors of performance on the particular three tasks in noise conditions due to individual differences in perceptual and cognitive abilities which are reflected in the variance in results. Importantly, since the study is primarily interested in performance across different maskers at the fixed signal-to-noise ratio, which was preserved throughout the trials, no tolerance threshold for individual maskers could be reported. Finally, while the results of this study show that speech perception and spoken word comprehension, both of which rely on working memory, are affected by noise maskers – it may be possible that auditory noise maskers also affect other cognitive processes (in non-auditory tasks) relying on working memory. Further research is recommended in order to find out more about this relationship.

6. References

- [1] J. Ou and S. Law, "Cognitive basis of individual differences in speech perception, production and representations: The role of domain general attentional switching," *Attention, Perception, & Psychophysics*, vol. 79, no. 3, pp. 945-963, Jan. 2017.
- [2] Z. G. Cai, R. A. Gilbert, M. H. Davis, M. G. Gaskell, L. Farrar, S. Adler and J. M. Rodd, "Accent modulates access to word meaning: Evidence for a speaker-model account of spoken word recognition," *Cognitive Psychology*, vol. 98, pp.73-101, Nov. 2017.
- [3] R. Smiljanić and D. Sladen, "Acoustic and semantic enhancements for children with cochlear implants," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 4, pp. 1085-1096, Jun. 2013.
- [4] X. Wang and L. Xu, "Speech perception in noise: Masking and unmasking," *Journal of Otology*, vol. 16, no. 2, pp. 109-119, Apr. 2021.
- [5] D. S. Brungart, B. D. Simpson, M. A. Ericson and K. R. Scott, "Informational and energetic masking effects in the perception of multiple simultaneous talkers," *The Journal of the Acoustical Society of America*, vol. 110, no. 5, pp. 2527-2538, Nov. 2001.
- [6] M. L. Lecumberri, M. Cooke and A. Cutler, "Non-native speech perception in adverse conditions: A review," *Speech Communication*, vol. 52, no. 11-12, pp. 864-886, Nov. 2010.
- [7] P. Assmann, and Q. Summerfield, "The Perception of Speech Under Adverse Condition," in Greenberg, S., Ainsworth, W.A. and Fay, R.R. (eds.) *Speech Processing in the Auditory System*. New York: Springer, pp. 231-308, 2006.
- [8] M. Imhof, "Listening Span Tests," in Worthington, D. L. and Bodie, G. D. (eds.) *The Sourcebook of Listening Research: Methodology and Measures*. Hoboken: John Wiley & Sons, pp. 394-401, 2019.
- [9] S. Kadota, *Shadowing as a practice in second language acquisition: Connecting inputs and outputs*. Milton Park: Routledge, 2019.
- [10] G. Underwood and N. Moray, "Shadowing and monitoring for selective attention," *Quarterly Journal of Experimental Psychology*, vol. 23, no. 3, 284-295, Aug. 1971.
- [11] S. Sumarsih, "The impact of shadowing technique on tertiary EFL learners' listening skill achievements," *International Journal of English Linguistics*, vol. 7, no. 5, 184-189, Jul. 2017.
- [12] P. Cullen, *Cambridge Vocabulary for IELTS Advanced Band 6.5+ with Answers and Audio CD*. Cambridge: Cambridge University Press, 2012.
- [13] BBC sound effects. (n.d.). Retrieved from <https://sound-effects.bbcrewind.co.uk>
- [14] R. Smiljanić and A. R. Bradlow, "Temporal organization of English clear and conversational speech," *The Journal of the Acoustical Society of America*, vol. 124, no.5, pp. 3171-3182, Nov. 2008.
- [15] S. V. Van der Feest, C. P. Blanco and R. Smiljanic, "Influence of speaking style adaptations and semantic context on the time course of word recognition in quiet and in noise," *Journal of Phonetics*, vol. 73, pp. 158-177, Jan. 2019.
- [16] L. Labia, L. Shtrepi and A. Astolfi, "Improved room acoustics quality in meeting rooms: Investigation on the optimal configurations of sound-absorptive and sound-diffusive panels," *Acoustics*, vol. 2, no. 3, pp. 451-473, Jun. 2020.
- [17] S. Lambert, "Aptitude testing for simultaneous interpretation at the University of Ottawa," *Meta*, vol. 36, no. 4, pp. 586-594, Dec. 1991.
- [18] A. Conway, M. Kane, M. F. Bunting, D. Zach Hambrick, O. Wilhelm and R.W.Engle, "Working memory span tasks: A methodological review and user's guide," *Psychonomic Bulletin & Review*, vol. 12, no. 5, pp. 769-786, Oct. 2005.
- [19] Y. Hu, Q. Tahmina, C. Runge D. R. Friedland, "The perception of telephone-processed speech by combined electric and acoustic stimulation," *Trends in Amplification*, vol. 17, no. 3, pp. 189-196, Nov. 2013.
- [20] E. M. Picou, J. Gordon and T. A. Ricketts, "The effects of noise and reverberation on listening effort in adults with normal hearing," *Ear & Hearing*, vol. 37, no. 1, pp. 1-13, Jan-Feb. 2016.
- [21] M. Rudner, V. Lyberg-Åhländer, J. Brännström, J. Nirme, M. K. Pichora-Fuller and B. Sahlén, "Listening comprehension and listening effort in the primary school classroom," *Frontiers in Psychology*, vol. 9, pp.1-7, Jul. 2018.
- [22] B. C. Moore and C. Tan, "Perceived naturalness of spectrally distorted speech and music," *The Journal of the Acoustical Society of America*, vol. 114, no. 1, pp. 408-419, Jul. 2003.
- [23] B. B. Monson, E. J. Hunter, A. Lotto and B. H. Story, "The perceptual significance of high-frequency energy in the human voice," *Frontiers in Psychology*, vol. 5, pp. 1-11, Jun. 2014.
- [24] M. Klatté, K. Bergström and T. Lachmann, "Does noise affect learning? A short review on noise effects on cognitive performance in children," *Frontiers in Psychology*, vol. 4, pp. 1-6, Aug. 2013.
- [25] N. Prodi, and C. Visentin, "On the relationship between a short-term objective metric and listening efficiency data for different noise types," *The Journal of the Acoustical Society of America*, vol.141, no.5, pp. 3972-3972, May. 2017.
- [26] S. L. Mattys, M. Davis, A. R. Bradlow and S. K. Scott, "Speech recognition in adverse conditions: A review," *Language and Cognitive Processes*, vol. 27, no. 7-8, pp. 953-978, Jul. 2012.
- [27] J.S. Snyder, M. K. Gregg, D. M. Weintraub and C. Alain, "Attention, awareness, and the perception of auditory scenes," *Frontiers in Psychology*, vol.3, pp. 1-17, Feb. 2012.