Perceptual evaluation of dissimilarity between auditory stimuli: an alternative to the paired comparison

Pierre-Yohan Michaud, Sabine Meunier, Philippe Herzog, Mathieu Lavandier, Gérard Drouet d'Aubigny

Acta Acustica united with Acustica
Vol. 99 (2013) 806 – 815

DOI 10.3813/AAA.918658

Perceptual Evaluation of Dissimilarity Between Auditory Stimuli: An Alternative to the Paired Comparison

Pierre-Yohan Michaud[1)], Sabine Meunier[1)], Philippe Herzog[1)], Mathieu Lavandier[2)], Gérard Drouet d'Aubigny[3)]

[1)] LMA, C.N.R.S., UPR 7051, Aix-Marseille Université, Centrale Marseille, 31 Chemin Joseph Aiguier, 13402 Marseille Cedex 20, France. michaud@lma.cnrs-mrs.fr

[2)] Université de Lyon, Ecole Nationale des Travaux Publics de l'Etat, Laboratoire Génie Civil et Bâtiment, Rue M. Audin, 69518 Vaulx-en-Velin Cedex, France

[3)] Laboratoire Jean Kuntzman (C.N.R.S.), Département de Statistique, 1251 avenue centrale B.P. 47, 38040 Grenoble Cedex 9, France

Summary

In order to examine the perceptual dimensions used by listeners to differentiate sounds, dissimilarity judgments are analyzed with a multidimensional scaling technique. The paired comparison method is often recommended as the standard method for collecting judgments of dissimilarity between audio stimuli. An alternative method is presented here. On each trial, listeners are asked to select which stimulus among three stimuli is the most similar to a reference stimulus. The method was tested wuth simulations and an actual listening test. Across trials, every stimulus was used in turn as the reference stimulus. The analysis of simulated data indicated that, the number of stimuli did not influence the estimation of dissimilarity, and that 20 simulated listeners were sufficient for recovering continuous dissimilarity values. The listening test conducted to evaluate the musical restitution by 12 loudspeakers led to two perceptual dimensions similar to those obtained in a paired comparison experiment.

PACS no. 43.66.Lj

1. Introduction

The goal of many studies on sound dissimilarity is to determine the perceptual dimensions used by listeners to differentiate the stimuli. These dimensions are often estimated based on the method of multidimensional scaling (MDS), which yields the estimated dissimilarity between the stimuli in a multidimensional space.

The paired comparison (PC) method is often considered as the method of choice for the study of sound dissimilarity [1, 2]. Accordingly, listeners evaluate the overall dissimilarity between paired stimuli, for example by giving an estimate on a scale ranging from "very similar" to "very dissimilar". An individual dissimilarity matrix is filled with listeners estimates. The individual dissimilarity matrices or their average can be analyzed with the MDS technique. The PC method has been used in several experiments, generally to evaluate the underlying dimensions of different perceptual attributes of sound sets (e.g. musical timbre [3], sound reproducing systems [4, 5], radiating bars and plates [6, 7, 8, 9] and cochlear implants [10]).

Received 5 March 2012,
accepted 12 June 2013.

A drawback of the paired comparison method is the number of trials which increases with the number of stimuli. Because of this limitation, several previous studies investigated alternatives to the paired comparison method [11, 12, 13, 14, 15]. Many of the alternative methods are based on an indirect estimation of dissimilarity. They can be classified in three categories based on the task carried out by participants: sorting, ranking or picking task.

The most commonly used alternative method to paired comparison is free sorting [15, 16]. Accordingly, participants are instructed to create groups of similar stimuli. The hierarchical sorting method is an alternative sorting method. On each step of this procedure, participants merge stimuli or groups of stimuli until all of the stimuli are merged in one group. The hierarchical sorting can start from the condition where each stimulus is in a separate group (complete hierarchical sort), from the groups resulting from a free sorting step [11], or from the groups resulting from a constrained sorting step where the number of groups of stimuli is decided by the experimenter (truncated hierarchical sorting [15]).

Another type of method is the conditional ranking method, which consists in ranking the stimuli from the most similar to the least similar compared to a reference

stimulus. Each stimulus in turn serves as the reference [11, 13].

The method of triadic combinations belongs to the picking task category. On each trial three stimuli are presented and participants are asked to pick the two most similar and the two most dissimilar stimuli within the triad [17]. A different picking task is called pick any k/n and consists in choosing those k stimuli among the presented n stimuli that are more similar to a given reference [11]. Similarly to the conditional ranking method, each stimulus is, in turn, considered as the reference.

Sorting and ranking methods have been adopted in a number of previous auditory perception experiments. For example, the free sorting task was employed and compared to the PC method by Bonebright [16], Parizet and Koehl [18], Giordano *et al.* [15]. The hierarchical sorting task was also tested by Giordano *et al.* [15] and the triadic combination method was used by Novello *et al.* [19] in a musical similarity experiment.

To our knowledge, the pick any method presented by Rao and Katz [11] and Bijmolt and Wedel [14] has never been applied and studied for the evaluation of sounds. Moreover, it has never been compared to the standard paired comparison method. The only practical example that we found in the literature of an application of such method is a visual experiment conducted by Rogowitz *et al.* [20]. They ran an experiment using the pick any k/n method with the parameters $k = 1$ and $n = 8$. The term "pick any k/n" does not denote two key aspects of this method: the forced choice similarity picking task and the comparison of the presented stimuli to a reference, with each stimulus used in turn as the reference. For these reasons, in the following we term this method "Similarity Picking with Permutation of References" (SPPR).

The aim of this paper is to present the SPPR method as an alternative method for the evaluation of sound dissimilarity. The first part of this paper details the proposed method and its adaptation to listening tests. The second part is dedicated to the exploration of the method by simulating the judgments of listeners. The simulation principle and the resulting information concerning the effects of the number of stimuli and the number of listeners on the estimation of dissimilarity are presented. The last part of this paper details the comparison between the results of two listening tests involving the PC method and the SPPR method on the same set of 12 musical excerpts. The specific features of the SPPR method are discussed.

2. Presentation of the method

The data collection process proposed here is based on the method presented by Rao and Katz [11]. It was employed in the experiment conducted by Rogowitz *et al.* [20] to quantify with an MDS analysis the dimensions underlying the perceived dissimilarity of images. This last study was used here as a practical example to design a listening test based on the same method. On each trial of their experiment, 9 stimuli were selected randomly from a set of 97 and presented to the participant. One of these nine images was treated as a reference stimulus and the participant had to choose which of the eight remaining comparison images appeared most similar to this reference. The experiment corresponds to a pick $1/8$ task for each trial. Every image was used once as the reference and compared to the 96 other images randomly distributed into groups of 8 images without replacement. For a total of 97 stimuli, 12 presentations of 8 images were made for each reference. The entire test appears to be incomplete since every combination to draw 8 out of 96 images for each reference are not presented. This incomplete design seems convenient to reduce the number of trials to be evaluated by each participant. Based on the incomplete pick $1/p$ experiment design (p corresponding to the number of comparison stimuli on each trial), the method was adapted to auditory stimuli evaluation.

2.1. Test design for auditory stimuli evaluation

When adapting this method to the evaluation of auditory stimuli, some precautions should be taken. In the experiment designed by Rogowitz *et al.* [20], the participants could directly see the reference image along with the eight images to be compared. Comparison and reference stimuli could be compared simultaneously. Sounds cannot be compared simultaneously: the stimuli are therefore presented sequentially and participants have to rely more extensively on mnemonic resources in order to carry out the comparison. For example, listeners hear the reference first and then a set of other sounds to be compared to the reference, one after another. The strain on mnemonic resources is particularly daunting when relatively long sounds are involved (e.g., musical excerpts). Thus, it seemed mandatory to reduce the number of stimuli used in each trial, in order to adapt the comparison task from visual to auditory evaluation and to reduce the duration of each trial.

When evaluating a set of n sounds, each reference is compared to the other $(n-1)$ comparison sounds. Those $(n-1)$ sounds are randomly distributed in $(n-1)/p$ groups without replacement where p corresponds to the number of comparison sounds evaluated during a trial. Each sound is used as the reference, leading to an entire test of $n(n-1)/p$ trials. The number p has to be carefully chosen in order to have the best compromise between the number of sounds to be compared during one trial and the total number of trials for the entire test.

Table I presents the influence of the chosen number p on the number of trials needed for each reference and on the total number of trials, for the evaluation of a set of 40 stimuli. As a comparison, the evaluation of 40 stimuli with the PC method leads to an entire test of 780 trials. Table I shows that the number of trials rapidly decreases with the number of comparison stimuli p presented on each trial.

In order to adapt the task from a visual to an auditory evaluation, we chose to compare each reference sound to three comparison sounds ($p = 3$). Three comparison stimuli seemed to be the best compromise between the feasibility of the task for each trial and the total duration of

Table I. Number of trials presented for each reference, $(n-1)/p$, and number of trials in the entire test, $n(n-1)/p$, as a function of the number of comparison stimuli p presented during one trial for a set of 40 stimuli.

p	$(n-1)/p$	$n(n-1)/p$
2	19	780
3	13	520
4	9	360
5	7	280
6	6	240

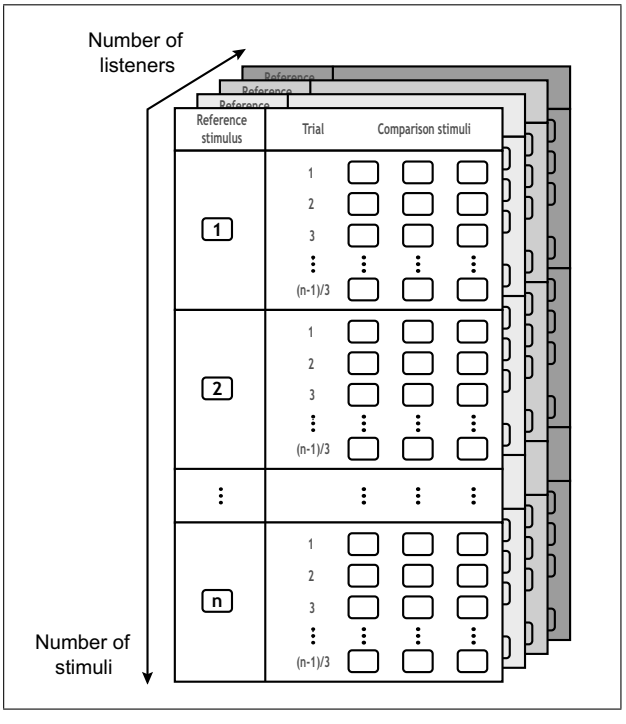

Figure 1. Design of an experiment involving the method of similarity picking with permutation of references. Each rectangle represents a sound. The comparison sounds presented with each reference are randomly distributed in triplets for each reference.

the entire test. The design of the SPPR method with 3 comparison sounds is presented in Figure 1. For each reference, the $(n-1)$ remaining stimuli are randomly distributed in $(n-1)/3$ triplets without replacement. This incomplete design of the SPPR method involves the evaluation of every stimulus as a reference, leading to an entire test of $n(n-1)/3$ trials. Note that conducting a complete test would involve the evaluation of every possible triplets compared to each reference and would lead to a very demanding experiment.

The dissimilarity evaluation is based on a forced choice task on each trial and does not imply any direct dissimilarity estimation. The participants have to choose which of the three comparison stimuli seems the most similar to the reference stimulus.

2.2. Accessing an average dissimilarity matrix

As was done by Rogowitz *et al.* [20], each individual matrix is first initialized to one and each diagonal is set to zero since the dissimilarity value between two identical stimuli is assumed to be null. The matrix is then modified depending on the listener judgments. During the experiment, when a stimulus j is judged within the triplet of comparison stimuli as the most similar to the reference i, the ij-th element of the individual matrix d_{ij} is decremented by one.

Judgments reported in each individual matrix may differ depending on whether i or j is presented as the reference. As a matter of fact, a pair (i, j) may appear in three kinds of presentation. Either i and j are comparison stimuli (not the reference) and play a similar role, or one is the reference stimulus and the other is one of the comparison stimuli, or the opposite. These last two cases could lead to asymmetries in the dissimilarity matrix since the choice of the comparison sound which is the most similar to the reference depends on the two other comparison sounds presented in the triplet. For example, an asymmetric result could occur when j is judged similar to the reference i in a trial, whereas i is not chosen as the most similar to the reference j in another trial. This example leads to different values of the symmetric elements: $d_{ij} \neq d_{ji}$. The MDS technique employed by Rogowitz *et al.* [20] requires a symmetric dissimilarity matrix to analyze the judgments of participants. Therefore it was proposed to keep the symmetric part of the matrix: $D_{ij} = (d_{ij} + d_{ji})/2$. All dissimilarity matrices appearing in this document have been processed following the same procedure.

At the end of the experiment and after the symmetrization step, every individual matrix cell can equal 1 if, for the pair (i, j), stimulus j is not judged as similar to the reference stimulus i and reciprocally, 0.5 if stimulus j is judged the most similar to the reference stimulus i but stimulus i is not judged the most similar to the reference stimulus j, and 0 if stimulus j is judged the most similar to the reference stimulus i and stimulus i is also judged similar to the reference stimulus j. Since the method consists in evaluating the $(n-1)/3$ triplets for each reference stimulus, judgments from each participant lead to a sparse matrix filled with 0, 0.5 or 1. Moreover, depending on the random distribution of the triplets for each reference, the test design differs from a listener to another. As each individual matrix is partially and differently filled by listeners answers, we chose to consider only the average of these matrices. This average matrix is then analyzed using an adequate MDS technique [11, 20].

3. Simulations

Simulating the judgments of a sample of listeners allows us to study the suitability of the SPPR method to estimate dissimilarities, evaluating the influence on the dissimilarity evaluation of the number of stimuli and simulated listeners. More specifically, simulations allow the reliability of the method and its potential biases to be evaluated.

Rao and Katz [11] conducted simulations on another adaptation of the "pick any" method. For a set of n stimuli, each trial included one reference and the $n-1$ comparison stimuli. The number of comparison stimuli to be picked as the most similar to the reference was set to 4 or 8. The procedure simulated by Rao and Katz [11] is thus not strictly the same as the procedure we have tested but the simulation presented in this section is based on a similar principle.

3.1. Principle

The dissimilarity judgments of the simulated listeners are computed from pre-determined dissimilarity values. These known dissimilarities were obtained from distance values, as follows: a random set of points was created in a space for which the number of dimensions was arbitrarily fixed to 3. The space created was randomly filled along the three dimensions. Then, the distance between each pair of points was considered as a dissimilarity value, and stored into a matrix. This resulting matrix was then considered as the reference, assumed to be the result of an ideal test. In the rest of the document, this original matrix will be called the true dissimilarity matrix.

These true dissimilarity values were used during the simulated trials to decide which of the comparison sounds would be chosen as most similar to the reference. A, B, C and D were the four stimuli involved in a simulated trial, A being the reference, and B, C and D the comparison stimuli. d_{AB}, d_{AC} and d_{AD} measure the true dissimilarity between A and B, A and C, A and D, respectively. The lowest value between d_{AB}, d_{AC} and d_{AD} corresponds to the highest similarity and indicates which stimulus among B, C and D would be chosen by the simulated listener as the most similar to the reference A in this simulated trial.

By simulating the method based on a random spatial configuration, Rao and Katz [11] were interested in measuring the correlation between the original space and the estimated space obtained with the MDS analysis of the estimated dissimilarity matrix. By simulating the judgments of listeners, our approach consists in comparing the estimated dissimilarity matrix resulting from the simulated test and the original true dissimilarity matrix, thus allowing the evaluation of the SPPR method. The transformation of the true dissimilarities into the estimated dissimilarities is the central point of the simulation. Two simulations were conducted to evaluate the influence of the number of stimuli and listeners on this transformation.

3.2. Design

3.2.1. Simulation 1: influence of the number of stimuli

Simulation 1 consisted in testing every possible set of three stimuli which could be compared to each reference during a complete experiment. The presentation of every possible triplet is easy to simulate but would be very demanding for an actual experiment. This simulation gathers the maximum of the dissimilarity information to fill the dissimilarity matrix. By presenting every combination of three out of $n-1$ stimuli, that is $(n-1)!/(3!((n-1)-3)!)$ combinations, the method is able to provide all the available information about the dissimilarity, as would do a single listener also evaluating every triplet. We used this simulation to assess the reliability of the method for a number of stimuli ranging from 10 to 50. In each case, for a fixed number of stimuli in the test, the correlation between true and estimated dissimilarities was calculated.

3.2.2. Simulation 2: influence of the number of listeners

Simulation 2 allowed us to simulate listening tests corresponding to more realistic scenarios. As described previously, one out of n stimuli was used as the reference, and compared to the other $(n-1)$ stimuli randomly distributed in the $(n-1)/3$ triplets. Hence, each listener does not evaluate every possible triplets. To simulate a full SPPR test, several listeners are then needed to obtain an average dissimilarity matrix. The aim of Simulation 2 was to determine the number of listeners required to obtain an average dissimilarity matrix equivalent to the one obtained in Simulation 1. For this simulation, the number of stimuli was arbitrarily fixed at 40. Thus, the entire test consisted of $n(n-1)/3$ trials which led to 520 trials for Simulation 2 compared to the 365560 trials needed for Simulation 1 with 40 stimuli. For each reference, the 39 comparison stimuli were distributed in 13 triplets.

First Simulation 2 was conducted as described. Then, simulations corresponding to more realistic experiments were also conducted by adding a random noise to the simulated listener responses. This investigates the robustness of the method with simulated listeners giving less-ideal answers. The listener-specific noise was added to each trial. In every trial, each of the three true dissimilarity values was modified by adding a randomly selected independent number (different for each dissimilarity value). These number was drawn from a Gaussian random distribution centered on the true dissimilarity to be modified. Three different noisy simulations were conducted with three noise distributions which variance values were set to 0.0016, 0.008 and 0.02. To evaluate the influence of the noise, the ratio of modified responses over the number of total responses was measured to get a simple figure of how many responses had been modified due to the added noise. The three distributions led to three modification ratios: 10, 20 and 30%.

3.3. Results

To evaluate the suitability of the SPPR method, correlations were calculated between the matrices of true and estimated dissimilarities. In order to employ the appropriate correlation coefficient, the relationship between true and estimated dissimilarities was evaluated. Figure 2 shows the estimated dissimilarity data, resulting from Simulation 1 with a number of stimuli arbitrarily set to 40, as a function of the true dissimilarity data. The histograms representing the distribution between 0 and 1 of the true and estimated dissimilarities are presented on the same

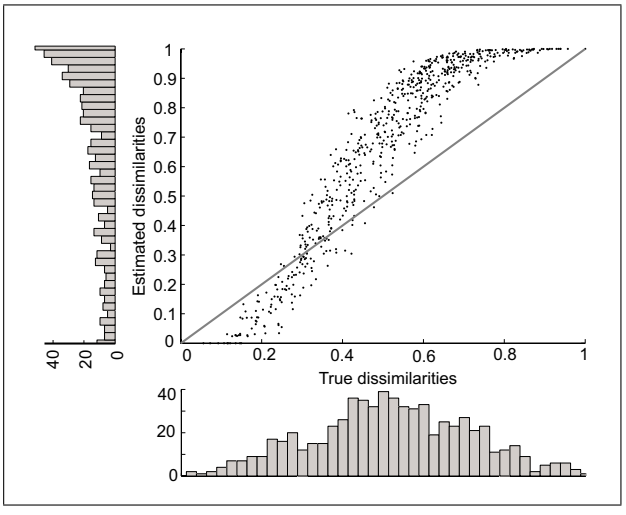

Figure 2. Non-linear relation between true and estimated dissimilarity values obtained with Simulation 1 for 40 stimuli and histograms presenting their distributions between 0 and 1.

figure. These histograms will be described later. The relation between the true and estimated dissimilarity values is clearly non-linear, and so the linear Pearson coefficient is not appropriate to evaluate this relation. Therefore, for both simulations, the Spearman correlation coefficient was preferred to quantify the link between dissimilarities. This coefficient offers the advantage of being robust to nonlinear relationships as it deals with the ranks of each element included in both matrices to evaluate the monotonicity of their relation (the value of the Spearman correlation corresponds to the Pearson coefficient calculated based on the ranks of both data sets). Correlations were calculated on the set of values corresponding to half of each symmetric matrix without taking into account the diagonal which was not modified during the test.

3.3.1. Number of stimuli

Considering the results of Simulation 1, the Spearman correlation coefficient was measured for a number of stimuli varying between 10 and 50. It remained almost constant with an average value of $\rho = 0.96$ and a standard deviation of 0.01 (not plotted as the correlation is constant). Simulation 1 thus shows that the number of stimuli does not influence significantly the correlation between the true and estimated dissimilarity matrices when every possible triplet presented to each reference is simulated. However, the simulation of the method does not allow the dissimilarity values to be recovered completely. This bias will be discussed later.

3.3.2. Number of listeners

In order to investigate the differences between Simulations 1 and 2, the Spearman correlation coefficient between the estimated dissimilarity matrices resulting from each simulation was calculated, as a function of the number of listeners involved in Simulation 2, for 40 stimuli. When the number of simulated listeners increased, the dissimilarity value resulting from Simulation 2 converged to the value obtained in Simulation 1. For a sample size of 20 or more simulated listeners, the method led to the same information that would be obtained if all the possible triplets were presented for each of the reference stimuli. For 20 simulated listeners, the correlation coefficient between dissimilarity matrices resulting from Simulations 1 and 2 already reached $\rho = 0.98$ (although the total number of trials is only $20x520 = 10400$, far less than the 365560 trials required in Simulation 1).

The method of similarity picking with permutation of references applied to a set of n stimuli tested only $(n-1)/3$ out of $(n-1)$ trials for each reference for Simulation 2. The results obtained when comparing Simulation 2 with Simulation 1 show that the influence of this partial filling of the dissimilarity matrix is efficiently reduced by using about 20 simulated listeners.

In order to evaluate the performance of the SPPR method in Simulation 2, the Spearman correlation coefficient between the estimated and true dissimilarity values was computed for 40 stimuli, as a function of the number of simulated listeners. Figure 3 illustrates the evolution of the correlation from one listener to 100 listeners, when simulated listeners (without noise) and when more "noisy" listeners were simulated (with added noise leading to 10, 20 and 30% of modification of the simulated listeners' responses). Without noise, the plot shows a steep increase in correlation coefficient for samples of up to 20 simulated listeners. For samples exceeding 20 listeners, the curve reaches a stable value. When adding noise to the responses, the shape of the correlation curves is modified but only for small numbers of listeners. The more noise is added, the more correlations decrease. Note that for up to 20% of modification the correlation curve is only slightly reduced. Even for 30% of modified responses, the Spearman correlation is still above 0.90 for 15 simulated listeners. When reaching 100 simulated listeners, the effect of the added noise is marginal.

3.4. Bias characterization

For the simulations with and without noise, even for 100 simulated listeners, the Spearman correlation coefficient does not exceed $\rho = 0.96$ (Figure 3). This correlation value, also obtained in Simulation 1, indicates that the method is never able to recover completely the dissimilarity information.

To determine the origin of this partial convergence shown by the simulations, the transformation from true to estimated dissimilarity data was studied. First, the relation between both data was shown to be non linear and monotonically increasing (Figure 2). In order to understand how the method transforms the true dissimilarity into estimated data, the distributions of true and estimated dissimilarity values are also displayed in Figure 2.

The distribution of the true dissimilarities, resulting from a uniform random distribution of the stimuli along the space dimensions, appears to follow a Gaussian-like shape. The estimated dissimilarity histogram however shows a shift of the values toward larger values. Indeed,

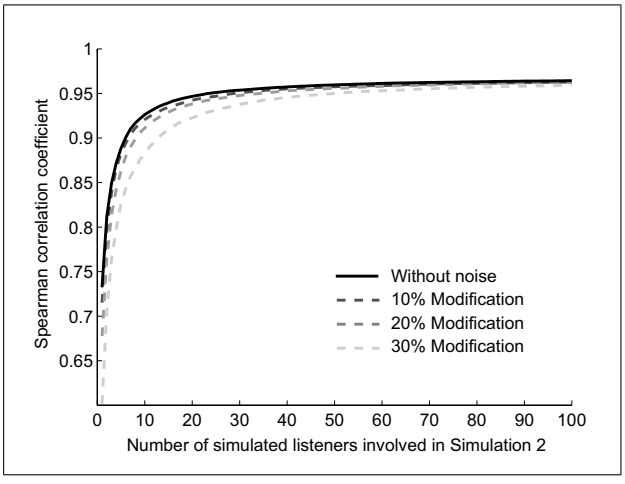

Figure 3. Correlation between true and estimated dissimilarity matrices in Simulation 2 as a function of the number of listeners for 40 stimuli. The original simulation and the noisy simulations are presented. The percentage of modification equals the proportion of modified responses given by the "noisy" simulated listeners over the total number of responses.

in Figure 2 most of the points are above the equality line. This shift towards larger dissimilarity values is due to the way the similarity information is coded. At each trial, a piece of information is obtained concerning the similarity between the reference stimulus and the chosen stimulus, but the dissimilarity value is not changed for the comparison stimuli which were not chosen, and therefore keeps its default value. For an entire test, before the symmetrization step, the dissimilarity matrix is filled with one third of listener responses with a null value and the remaining two thirds keep their initial value of 1.

To test if the true dissimilarity distribution could influence the simulation results, different distributions were tested. The true dissimilarity matrix obtained from the space was replaced by either a uniform distribution, an asymmetric Poisson distribution set between 0 and 1 or a Gaussian distribution centered on 0.5 with a variance of 0.17 and truncated between 0 and 1. Simulation 2 conducted with any of these three distributions led to an identical Spearman correlation coefficient of 0.98 between true and estimated dissimilarities for 100 simulated listeners. This correlation is the same for the three distributions and is just above the correlation obtained with the simulation based on the true dissimilarities resulting from an evenly filled space. Even with those three distributions, a small bias remains. There is still a systematic bias which leads to dissimilarity values that are overestimated compared to the true dissimilarity values for the different tested distributions. The histograms in Figure 2 confirms that the method overestimates the higher dissimilarities. However, this bias does not change the ranking of the averaged dissimilarity data, thus leading to a high Spearman correlation coefficient.

The possibility to simulate the SPPR method allowed us to estimate this bias which seems to originate from the dissimilarity construction rather than from the properties of the true dissimilarity matrix or from the noise added to the responses of simulated listeners. The bias is intrinsic to the design of the method but is relatively small. In order to assess its practical importance, the method of similarity picking with permutation of references was applied to an actual listening test.

4. Listening test: Application to the evaluation of dissimilarity between loudspeakers

The SPPR method was tested on an actual sound corpus. A listening test was conducted to evaluate the sound reproduction of loudspeaker recordings. The results were compared to those obtained by Lavandier *et al.* [4, 5], who used paired comparisons, the method mostly employed when evaluating loudspeakers [21, 22, 23, 24]. Here, the perceptual results obtained by Lavandier *et al.* and by the SPPR method are directly compared. First of all, the dissimilarity values resulting from the two methods are compared. Then, the MDS analysis of the dissimilarity data allows the comparison of the resulting perceptual spaces.

4.1. Stimuli

The work conducted by Lavandier *et al.* [4, 5] aimed at investigating the restitution of musical excerpts by several loudspeakers. The stimuli consisted of the sound radiated by 12 single loudspeakers recorded at a same position in the same room using the stereophonic technique AB-ORTF. The recordings were reproduced under headphones during the listening test. The musical excerpt was McCoy Tyner "Miss Bea" (3.3 s). All details concerning these recordings, loudspeakers, room and signal can be found in Lavandier *et al.* [4, 5].

4.2. Listeners

Twenty-seven normal-hearing participants took part in our listening test. This corresponds to the number of listeners involved in the PC experiment conducted by Lavandier *et al.* [4]. None of them had been trained in the perceptual evaluation of loudspeakers.

4.3. Procedure

For 12 loudspeakers, each reference sound was compared to the 11 other sounds, which were randomly distributed in four triplets for each listener. The 11 sounds did not fill the 4 triplets completely, therefore the last triplet was completed with one sound randomly chosen within the first 3 triplets. The entire test comprised 48 trials for each listener and was divided into two sessions which lasted about 10 minutes each, and were separated by a small break. As a comparison, the PC experiment conducted by Lavandier *et al.* [4] lasted about 20 minutes.

Listening tests were conducted with a graphical user interface that comprised four square buttons corresponding to the four stimuli. On each trial, the four stimuli (reference + three comparison sounds) were initially presented

in sequence. After this initial presentation, listeners could listen to any of the stimuli, as many times as they wanted, by clicking on the corresponding button. In order to conclude the trial, they had to choose the stimulus most similar to the reference. Listening tests were carried out in a soundproof booth and the stimuli were reproduced through headphones (Stax SR Lambda Professional; same model as in [4]).

4.4. Results and comparison

4.4.1. Perceptual dissimilarities

The dissimilarity data resulting from our experiment were compared to the data obtained by Lavandier *et al.* [4] with the PC method. For both methods, the diamonds on Figure 4 represent the Pearson correlation coefficients between the average matrix computed from all individual matrices, and the average matrices computed for subsets of individual matrices. The increasing number of listeners (x axis) corresponds to the size of each subset considered in the comparison. Each diamond represents the average of 100 Pearson correlation coefficients computed from random subsets of individuals of the same size. Results for both PC and SPPR methods show a very similar trend. The correlation coefficient asymptotes its maximum value for subsets containing at least 20 individual matrices. Twenty listeners appeared sufficient for recovering a reliable group-average dissimilarity matrix for both the PC and SPPR method.

Black dots on Figure 4 present the Spearman correlation coefficients between the average matrix resulting from the PC method, and the average matrix computed for subsets of individual matrices resulting from the SPPR method. Again, for each subset size (x axis), Figure 4 presents the average and standard deviation for 100 random subsets of listeners. The correlation increases sharply between 0 and 10 listeners and remains constant when reaching 20 listeners. This tendency is quite similar to what was observed for the simulations.

The maximum Spearman correlation coefficient between the SPPR and PC matrices is 0.75 (black dots on figure 4). This value indicates that the matrices resulting from the two listening tests are overall quite similar. Note that for this comparison, the mean dissimilarities are obtained from real (i.e. not simulated) listeners. This might explain why the correlation between the PC and SPPR results is lower than for the simulations.

Figure 5 summarizes an analysis of the reliability of the group-average dissimilarities obtained with the SPPR method based on a method similar to what used Giordano *et al.* [15]. Two groups of size N were picked randomly (here without replacement) within the 27 individual matrices and the average matrix of each group was computed. As an estimation of reliability, the correlation coefficient R^2 was calculated between the average matrices of each group. The average value over 10,000 draws was computed and presented for values of N between 1 and 13. This group-average data reliability analysis was done for both SPPR and PC methods.

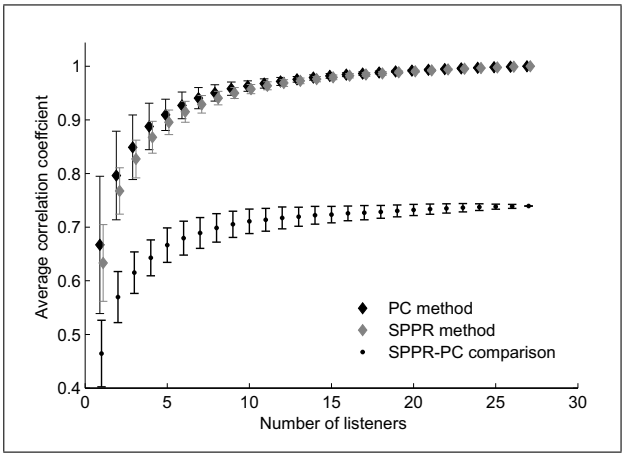

Figure 4. Pearson correlation coefficients between subsets of individual matrices and the group-average matrix, for PC method (black diamonds) and SPPR method (grey diamonds). Spearman correlation coefficients between subsets of individual SPPR matrices and the average PC matrix (black dots). Bars represent ± one standard deviation.

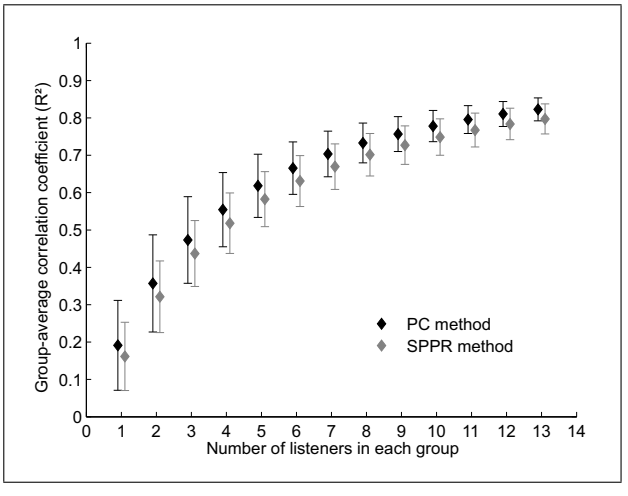

Figure 5. Group-average correlation coefficient (R^2) associated with its standard deviation as a function of the number of listeners in each group for both PC (black diamonds) and SPPR method (grey diamonds).

The reliability values obtained for the PC method are close to the ones presented in [15]. The reliability values obtained with the SPPR method appear to be very close to those obtained with the PC method, even closer than those obtained with the alternative methods tested in [15]. Note that for small values of N, the reliability might be overestimated because of the initial filling of the individual dissimilarity matrices. For our maximum group size, both methods have a reliability value close to 0.80. This is probably related to the experimental noise associated with the judgments by real listeners.

4.4.2. Perceptual MDS spaces

The dissimilarity data were analyzed using a metric MDS technique based on the SMACOF (Scaling by Majorizing a Complicated Function) algorithm [1], which was used

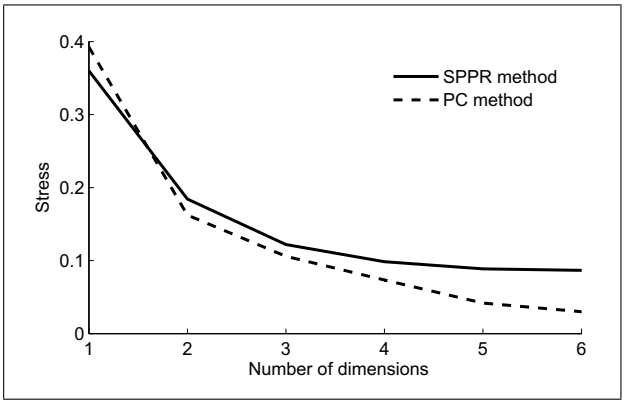

Figure 6. Stress plots for MDS analysis resulting from the PC method (dashed line) and from the SPPR method (plain line).

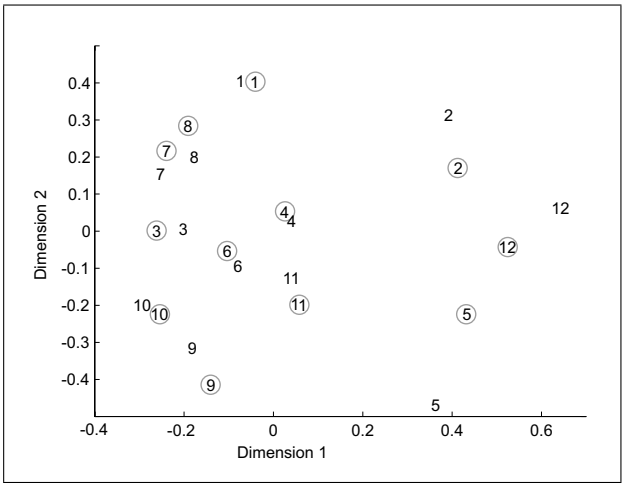

Figure 7. Comparison of the two-dimensional perceptual spaces obtained with the PC method and the SPPR method (circled numbers).The numbers correspond to each loudspeaker recording.

by Lavandier *et al.* [5]. Figure 6 shows the stress plot for both methods as a function of the number of dimensions. Two dimensions were retained for both experiments. Any spatial configuration resulting from the SMACOF algorithm can be stretched or rotated. In order to compare the two MDS spaces, a Procrustes adjustment procedure that rotates a configuration to match another configuration was performed. As a visual comparison, Figure 7 presents the auditory spaces derived from the dissimilarity matrices obtained with the PC and SPPR methods. It can be observed that the two methods provided very similar two-dimensional spaces: the positions of the loudspeakers are very similar in the two perceptual spaces.

To compare the MDS spaces, the Pearson correlation between the coordinates of the points representing the stimuli was calculated for each dimension. This correlation coefficient equals 0.99 for the first dimension and 0.91 for the second dimension. During the recording session, Lavandier *et al.* [4] recorded the reproduction of two recordings from two identical loudspeakers (loudspeakers 7 and 8). As expected, those two loudspeakers appeared very close in the MDS space resulting from the PC experiment. This is also true for the SPPR experiment. For further information about the dimensions and their interpretation, the reader is referred to Lavandier *et al.* [5].

5. Discussion

5.1. Validity of the SPPR method

The results of the simulations together with the analysis of the actual listening test indicated that the SPPR method seems able to recover an average dissimilarity matrix close to the one that can be obtained with the more commonly adopted PC method. Even if the SPPR method is characterized by a systematic bias (see section 3.4), it does not impact much the perceptual space revealed with an MDS analysis of the data. Conversely, the SPPR method does not provide individual dissimilarity matrices: their content is only a rough estimate of the individual-level dissimilarity, because of the effects of the structure of the picking task itself. Interestingly, averaging over a limited number of listeners nonetheless allows to get an estimate of the underlying dissimilarity which seems reliable enough to allow the modeling of the underlying perceptual dimensions. The need of only a limited number is again confirmed by both our simulations and the listening test: around 20 listeners seemed to be sufficient. For the listening test, this number was similar for both SPPR and PC method. Moreover, the reliability estimated for both methods was very close, and was characterized by a highly similar effect of the number of listeners. As far as the average dissimilarity matrix is concerned, the SPPR method seems a valid alternative to the PC method.

5.2. Duration of SPPR experiments

With the proposed SPPR method, each stimulus is used in turn as the reference, which leads to an entire test of $n(n-1)/3$ trials, n being the number of stimuli. During each trial, participants have to listen to a minimum of 4 sounds. As a comparison, $n(n-1)/2$ trials are needed in PC experiments, with a minimum of 2 sound playbacks for each trial. These are strict minima which do not take into account the potential need for listening several times to some excerpts - e.g. when dissimilarities are small or due to listener's fatigue. The minimum number of sounds to be listened to is 33% higher for the SPPR than for the PC method.

Assuming that the two methods are equally demanding, an experiment should last significantly longer for the SPPR method than for the PC method. Conversely, our experiment showed that both tests last about the same time (20 minutes for 12 stimuli). Such an overall duration is only a very coarse estimate, but this comparison may indicate that the task involved in the SPPR method was experienced by the listeners as simpler than a direct dissimilarity evaluation. Indeed, even with 4 sounds per trial, the need to listen again to some excerpts or the time needed to make a decision were reduced. This corresponds to some of the

listeners comments which were gathered at the end of each test. Unfortunately, we did not record event data sustaining this hypothesis when conducting our experiments. It has still to be backed by future work.

5.3. Further developments

The majority of previous studies on paired comparison of sound stimuli involved a maximum of 20 stimuli. Evaluating more stimuli potentially leads to a more homogeneous repartition of the stimuli along the perceptual dimensions and provides a better description of the perceptual space. It might also allow the experimenter to reveal more dimensions compared to the perceptual results obtained with smaller sets of sounds. However, the number of presentations (pairs) of the experiment increases dramatically with the number of stimuli. Direct similarity ratings involving a large number of stimuli then become too demanding, as previously stated [1]. Such an experiment is time consuming and the fatigue or the loss of attention of listeners becomes significant [25]. Most of the studies evaluating large sets of sounds point out the fact that the evaluation of large sets with the paired comparison would be "inefficient" [13], "cumbersome" [11], "undesirable" [26] or "limited" by the number of pairs to evaluate [16]. For most listening experiments, it is then recommended to have listeners taking breaks between listening sessions in order to avoid the inconvenience of a long duration test. A maximum duration of 30 minutes per session is often recommended [23]. To evaluate a large set of audio stimuli, the experimenter must then divide a paired comparison experiment into several sessions.

As far as we know, there are no studies that have compared a PC dissimilarity estimation conducted during one single session and the same experiment run in several sessions. Poulton [27] underlined the sequential or "transfer from previous judgments" bias which exists between the judgment made during one trial and the judgments made during previous trials. Poulton [28] specified that "the sequential contraction bias is likely to affect any series of judgments, unless each stimulus is judged deliberately against a standard stimulus". The permuted references and the forced choice task involved in the SPPR method might then contribute to reduce such "sequential bias" by limiting the tendency of listeners to change their evaluation scale between trials. Splitting the test into several sessions could then be less concerning. This is the topic of ongoing work concerning the SPPR method.

6. Conclusion

The similarity picking with permutation of references (SPPR) method was inspired by a visual experiment conducted to evaluate the similarity between images. In this paper, it was adapted to auditory evaluation in order to propose an alternative method to the paired comparison (PC) method.

Simulations were conducted in order to investigate the effect of the number of stimuli and listeners on the estimation of dissimilarity. It was shown that the number of stimuli involved in the simulation did not influence this estimation. The simulations also showed that most of the dissimilarity information was gathered with about 20 simulated listeners. A small bias appeared in the estimation of dissimilarity leading to overestimate the dissimilarity values. This bias was characterized and seemed to originate from the design of the experiment.

A listening test was conducted with 12 loudspeakers and compared to the PC method. Comparing dissimilarity data and MDS spaces revealed that the SPPR method gives results very similar to those obtained with the standard PC method. Even if the PC and the SPPR methods require a different task from the listener, the average dissimilarity judgments obtained with the SPPR method were well correlated with those obtained by Lavandier *et al.* [4] with the PC method. The analysis of the dissimilarity matrices led to two-dimensional spaces whose dimensions were very similar to the dimensions obtained by Lavandier *et al.* [5].

The SPPR method seems to involve a simpler task and its embedded reference could help to divide an entire test into several sessions. This is a motivation for further work, aiming at the evaluation of the performance of the SPPR method for larger sets of stimuli.

Acknowledgement

The authors are grateful to the listeners who took part in the listening test. The authors also thank Bruno Giordano and an anonymous reviewer for very constructive remarks and suggestions.

References

[1] I. Borg, P. Groenen: Modern multidimensional scaling, theory and applications, 2nd edition. Springer, New York, 2005.

[2] Y. Takane, S. Jung, Y. Oshima-Takane: Multidimensional scaling. – In: Handbook of quantitative methods in psychology. R. E. Millsap, A. Maydeu-Olivares (eds.). Sage Publications, London, 2009, 219–242.

[3] J. M. Grey: Multidimensional perceptual scaling of musical timbres. J. Acoust. Soc. Am. **61** (1977) 1270–1277.

[4] M. Lavandier, P. Herzog, S. Meunier: Comparative measurements of loudspeakers in a listening situation. J. Acoust. Soc. Am. **123** (2008) 77–87.

[5] M. Lavandier, S. Meunier, P. Herzog: Identification of some perceptual dimensions underlying loudspeaker dissimilarities. J. Acoust. Soc. Am. **123** (2008) 4186–4198.

[6] G. Canévet, D. Habault, S. Meunier, F. Demirdjian: Auditory perception of sounds radiated by a fluid-loaded vibrating plate excited by a transient point force. Acta Acustica united with Acustica **90** (2004) 181–193.

[7] J. Faure, C. Marquis-Favre: Perceptual assessment of the influence of structural parameters for a radiating plate. Acta Acustica united with Acustica **91** (2005) 77–90.

[8] S. McAdams, A. Chaigne, V. Roussarie: The psychomechanics of simulated sound sources: Material properties of impacted bars. J. Acoust. Soc. Am. **115** (2004) 1306–1320.

[9] S. McAdams, V. Roussarie, A. Chaigne, B. L. Giordano: The psychomechanics of simulated sound sources: Material properties of impacted thin plates. J. Acoust. Soc. Am. **128** (2010) 1401–1413.

[10] C. M. McKay, H. J. McDermott, G. M. Clark: The perceptual dimensions of single-electrode and nonsimultaneous dual-electrode stimuli in cochlear implantees. J. Acoust. Soc. Am. **99** (1996) 1079–1090.

[11] V. R. Rao, R. Katz: Alternative multidimensional scaling methods for large stimulus sets. Journal of Marketing Research **8** (1971) 488–494.

[12] M. Subkoviak, A. L. Roecks: A closer look at the accuracy of alternative data-collection methods for multidimensional scaling. Journal of Educational Measurement **13** (1976) 309–317.

[13] L. Tsogo, M. H. Masson, A. Bardot: Multidimensional scaling methods for many-object sets: A review. Journal Multivariate Behavioral Research **35** (2000) 307–319.

[14] T. H. A. Bijmolt, M. Wedel: The effects of alternative methods of collecting similarity data for multidimensional scaling. International Journal of Research in Marketing **12** (1995) 363–371.

[15] B. L. Giordano, C. Guastavino, E. Murphy, M. Ogg, B. K. Smith, S. McAdams: Comparison of methods for collecting and modeling dissimilarity data: Applications to complex sound stimuli. Multivariate Behavioral Research **46** (2011) 779–811.

[16] T. L. Bonebright: An investigation of data collection methods for auditory stimuli: Paired comparison versus a computer sorting task. Behavior Research Methods, Instruments and Computers **28** (1996) 275–278.

[17] Y. Takane: The method of triadic combinations: A new treatment and its applications. Behaviormetrika **11** (1982) 37–48.

[18] E. Parizet, V. Koehl: Application of free sorting tasks to sound quality experiments. Applied Acoustics **73** (2012) 61–65.

[19] A. Novello, M. F. McKinney, A. Kohlrausch: Perceptual evaluation of inter-song similarity in western popular music. Journal of New Music Research **40** (2011) 1–26.

[20] B. E. Rogowitz, T. Frese, J. R. Smith, C. A. Bouman, E. Kalin: Perceptual image similarity experiment. Proceedings of the SPIE Conference on Human Vision and Electronics Imaging, San Jose, CA, 1998, 576–590.

[21] S. P. Lipshitz, J. Vanderkooy: The great debat: subjective evaluation. J. Audio Eng. Soc. **29** (1981) 482–491.

[22] A. Gabrielson, B. Lindstrom: Perceived sound quality of high fidelity loudspeakers. J. Audio Eng. Soc. **33** (1985) 33–53.

[23] AES20-1996: AES recommanded practice for professional audio - subjective evaluation of loudspeakers. J. Audio Eng. Soc. **44** (1996) 383–401.

[24] IEC Publication 60268-13: Sound system equipment - Part 13: Listening tests on loudspeakers. International Electrotechnical Commission, Geneva, Switzerland, 1998.

[25] G. P. Scavone, S. Lakatos, C. R. Harbke: The sonic mapper: An interactive program for obtaining similarity ratings with auditory stimuli. Proceeding of International Conference on Auditory Display, Kyoto, Japan, 2002, 368–371.

[26] I. Spence, D. W. Domoney: Single subject incomplete designs for nonmetric multidimensional scaling. Psychometrika **39** (1974) 469–490.

[27] E. C. Poulton: Models of biases in judging sensory magnitude. Psychological Bulletin **86** (1979) 777–803.

[28] E. C. Poulton: Biases in quantitative judgements. Applied Ergononics **13** (1982) 31–42.