ARCHIVES OF ACOUSTICS
32, 3, 729–735 (2007)

NEW METHOD FOR THE ESTIMATION OF HOMOGENEITY OF INDIVIDUAL JUDGEMENTS IN PSYCHOACOUSTICAL EXPERIMENTS

Elżbieta ARANOWSKA, Maria MANTURZEWSKA, Piotr ROGOWSKI

The Fryderyk Chopin Academy of Music
Okólnik 2, 00-368 Warszawa, Poland
e-mail: pro@chopin.edu.pl

(*received May 21, 2007; accepted July 18, 2007*)

This paper presents a new indicator for the estimation of homogeneity of individual judgements in psychoacoustical experiments. The indicator may be also used in studies from other fields in which the judgements are obtained in the form of numerical values. Two examples of the application of the new indicator have been discussed: one for the assessment of cohesion of listeners' judgements in a psychoacoustical experiment conducted with the use of the method of absolute magnitude estimation and the other one for the assessment of homogeneity of marks given by jury members in a music competition. In the latter case, a new aggregate measurement tool was also proposed for ranking of competition participants.

Keywords: auditory judgements, statistical analysis.

1. Introduction

Individual measurements fulfil a significant role in many areas of acoustics – particularly in sound quality studies conducted in psychoacoustics, musical acoustics and electroacoustics. In the majority of cases, the judgements obtained in experimental studies have the form of numerical values or may be converted to such a form. A significant issue, both at the stage of preparing an experiment as well as during interpretation of its results is the assessment of the homogeneity of judgements. An over-dispersion of the data may indicate an incorrect selection of the group of listeners, for example, including persons who do not possess proper competences, incorrectly formulated auditory tasks or inappropriate arrangement of experimental conditions.

Homogeneity of data is an important issue in various kinds of auditory judgements. In many cases, such as sound quality assessment of electroacoustic equipment and evaluation of performances during musical competitions, the aim of auditory judgements is to rank the objects undergoing assessment. In such a case, the dispersion of judgements is usually neglected and the only criterion used for ranking is the mean value of judgements.

To date, statistics lacked a measure that would express the level of concordance of the judgements obtained from a group of jury members for a single object undergoing assessment. The most commonly used measure of agreement among raters is the Kendall's W coefficient of concordance which expresses the level of concordance for an entire group of judged objects. In this paper, presented is a new index for the assessment of homogeneity of judgements, λ. The possible applications of the new index are illustrated by two examples from the fields of psychoacoustics and music.

2. The λ judgement homogeneity index

The λ judgement homogeneity index was defined and described by ARANOWSKA [1]. The index is given by:

$$\lambda = 1 - \frac{\sigma_o^2}{\sigma_{\max}^2}, \tag{1}$$

where σ_o^2 – variance obtained, $\sigma_{\max}^2$ – maximum variance.

The value denoted as σ_o^2 is the actual variance obtained in the experiment whereas the maximum variance, $\sigma_{\max}^2$, refers to a hypothetical situation wherein the objects are judged according to contradictory criteria. The judgements obtained from jury members are then grouped at the extremes of the scale creating a variate with a two-point distribution with equal probability ($p = 0.5$) of assuming the minimal value or the maximal value. Unlike variance, the proposed λ index is therefore a normalised measure, referring to the deviation from maximum variance. When the variance of judgements increases towards maximum, the λ index approaches zero whereas in the case of full concordance of judgements λ reaches a value of 1. It is possible to determine the statistical significance of the λ indexes assigned to the individual objects (test tasks, stimuli, performers) as well as the significance of differences between pairs of λ indexes [cf. 1].

The λ' index, introduced in this study, has similar properties to the λ index:

$$\lambda' = 1 - \sqrt{\frac{\sigma_o^2}{\sigma_{\max}^2}}. \tag{2}$$

The advantage of the λ' index is that the differences between its values are larger and more pronounced in typical experimental conditions.

The λ and λ' indexes may be used to develop KA [cf. 1] and KA' aggregate measures. The KA' aggregate criterion, introduced in this paper, is given for an i-th test item by:

$$KA_i' = \frac{2}{\dfrac{1}{x_i'} + \dfrac{1}{\lambda_i'}} (x_i' \neq 0 \quad \text{and} \quad \lambda_i' \neq 0), \tag{3}$$

where:

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \tag{4}$$

and the values of $x_{\min}$ and $x_{\max}$ denote respectively the minimal and maximal marks on the scale. Aggregate measures make it possible to determine rankings of judged objects that take into account not only the means of judgements but also their consistency across jury members. Aggregate measures give preference to objects judged consistently, with high λ or λ' values. If the judgements are inconsistent (low λ or λ' values) the rank of an object is lowered.

3. Example applications of the λ' index

The possible applications of the λ' index are illustrated in this paper on two sets of data: the marks given by jury members during a music competition and the results of a psychoacoustical experiment. The intention of the authors is not to discuss the data in detail, but to demonstrate example applications of the λ' index and point at its specific analytical features.

The first set of data are the judgements made by jury members during a prestigious international music competition in Poland [2]. Seventy seven participants took part in the competition. The competition had three stages – 38 participants advanced to the second stage, and 12 to the finals. The jury comprised 28 prominent musicians who judged the performances on a scale from 0 to 25 points. The ranks of the participants were determined during the competition in a usual way, by calculating the arithmetic mean of the marks given by the jury members.

Figure 1 shows the λ' index values calculated in the present study for all participants in the successive stages of the competition. The data were ordered according to the ranks after the first stage of the competition. The λ' values calculated for the first stage have also been presented in Fig. 3 and compared with the actual ranks of the participants in the competition.

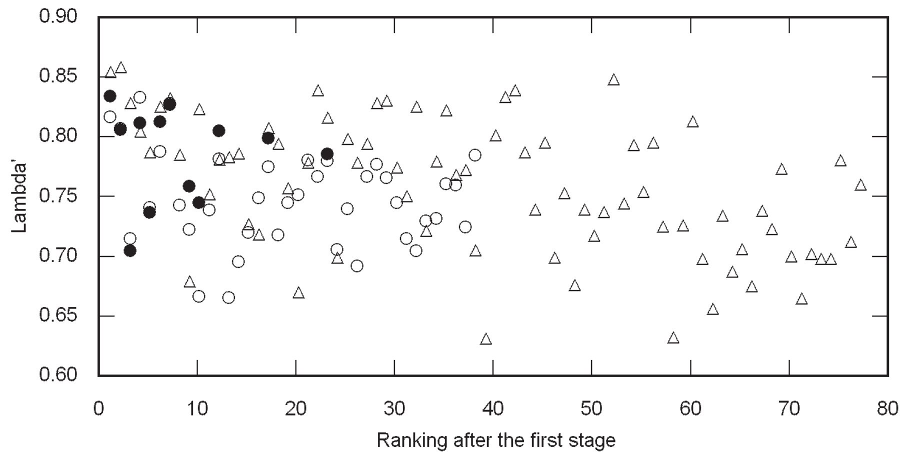

Fig. 1. Cohesion, represented by the λ' index values, of judgements given by the jury members at three stages of a music competition: △ – first stage of the competition, ○ – second stage results, ● – third stage.

Figure 2 shows an example application of the KA' aggregate measure. The data plotted give a comparison of the actual ranks of the participants after the three competition stages with the ranks determined on the basis of the KA' aggregate measure. Figure 3 shows the λ' values separately for the participants at the three stages of the competition. The data plotted in Fig. 3 indicate that the traditional method of qualification eliminated from the competition a number of participants who were judged quite

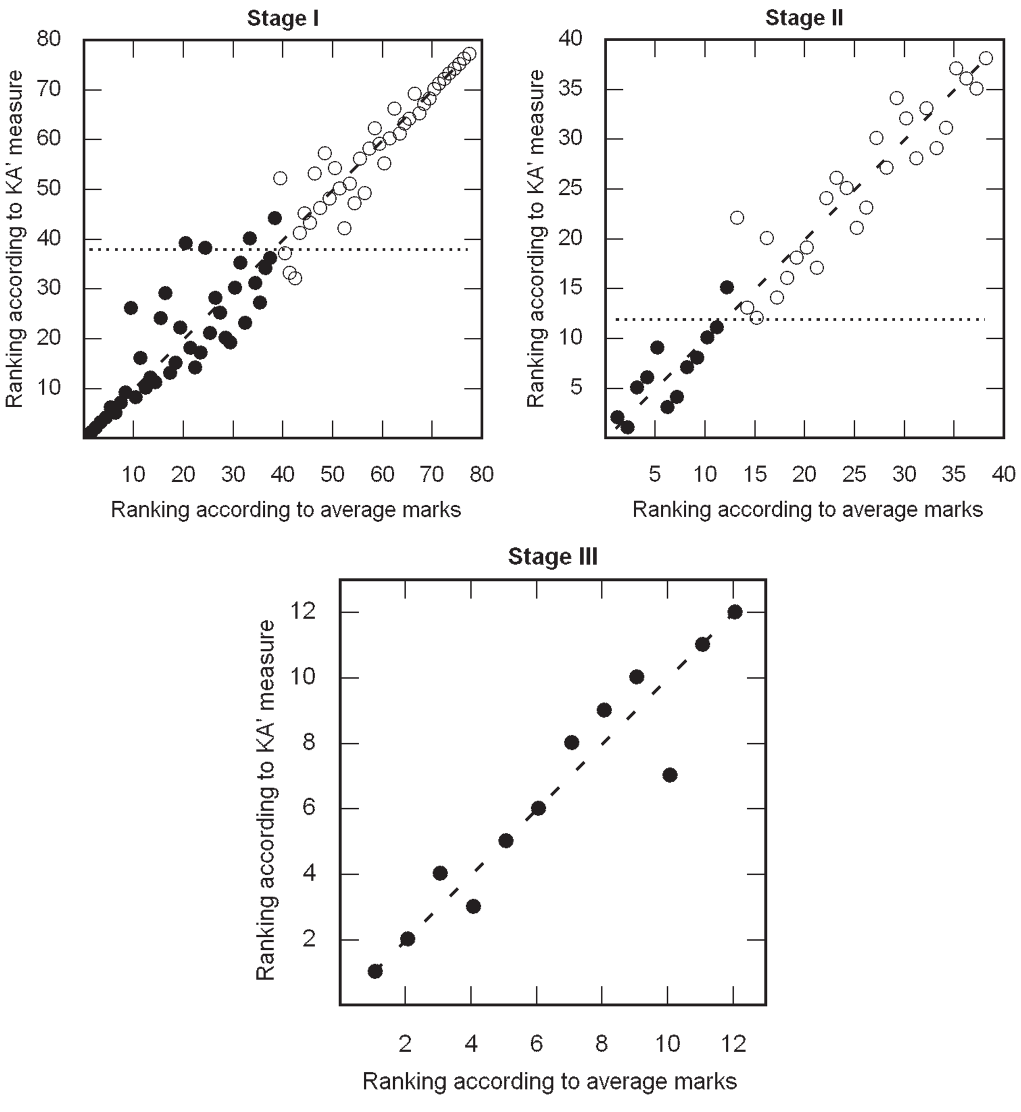

Fig. 2. Results of the three stages of a music competition – comparison of the actual ranks of the participants with ranks determined on the basis of the aggregate measure KA'. In the two first panels closed circles indicate the participants who advanced to the next stage and open circles refer to those who failed to qualify.

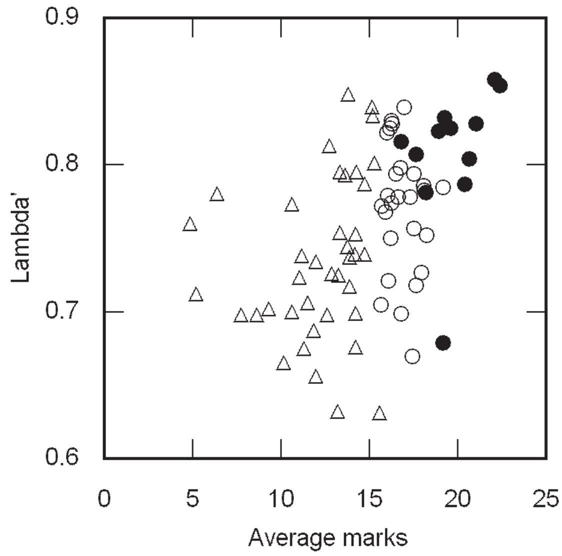

Fig. 3. Cohesion of the jury judgements represented by the λ' index values, plotted against the ranks of the participants after the first stage of a music competition. The different symbols indicate the future finalists of the competition (●), the participants of the second stage (○) and the participants who were eliminated after the first stage (△).

highly and coherently and favoured those who obtained only a slightly higher mean score, but the marks were incoherent. It should be noted that the aggregate measurement reduces the influence of overestimating or underestimating the performance of a participant by a jury member and – in contrast to so-called *weak statistics* – gives unequivocal results.

The second set of example data was taken from an experiment the purpose of which was to investigate the sensation roughness produced beating tone pairs [3]. A detailed description of the experimental procedure can be found in MIŚKIEWICZ *et al.* [3]. The data were obtained for seven series of tone pairs, each centred at a different frequency. The centre frequencies of the tone pairs, *i.e.* the geometric means of the two tone frequencies, were 63, 125, 250, 500, 1000, 2000 and 4000 Hz. The main experimental variable was the beat rate, corresponding to the frequency difference between the two tones in a pair. The lowest beat rate was 4 Hz and the highest rate ranged from 148 Hz for a 63-Hz centre frequency to 512 Hz for 4000 Hz. Seven experienced listeners took part in the experiment and their task was to assign a number to the sensation of roughness produced by each stimulus in a series of tone pairs centred at a given frequency. The listeners were instructed to use only positive numbers in their judgements. Each listener made 10 judgements of each of the stimuli. As an example of application of the λ' index, in Figs. 4 and 5 shown are the data obtained for a centre frequency of 4 kHz.

Figure 4 shows the cohesion of judgements obtained from the group of listeners in successive measurement series. Solid line indicates the λ' index values calculated for the full set of data. To demonstrate the variability of λ' values calculated for different

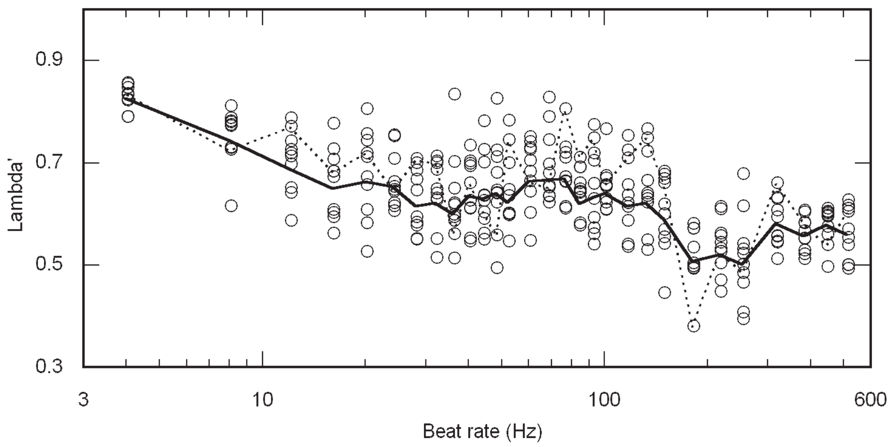

Fig. 4. Cohesion of results in 10 measurement series. Solid line indicates the λ' index values calculated for the full set of data. To demonstrate the variability of λ' values calculated for different stimuli, the dotted line shows the λ' index values for one, selected measurement series.

stimuli, the dotted line shows the λ' index values for one, selected measurement series. Large differences in the λ' index value indicate large variability of judgements, possibly caused by inconsistent judgemental criteria or by contextual effects.

Figure 5 shows the λ' index values determined for the data obtained in 10 series of judgements from each individual listeners. The discrepancies between individual λ' values and the λ' index determined for the whole group (Fig. 5, solid line) indicate an inconsistency among listeners in their judgements of certain stimuli, an effect that has been observed in an analysis of group data and those obtained from the listener whose judgements were most cohesive.

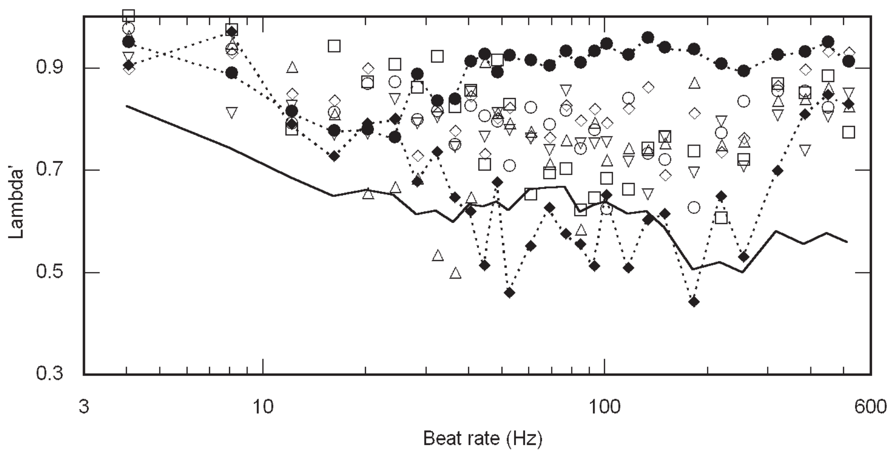

Fig. 5. Cohesion of results obtained from each of the seven listeners who participated in the experiment. The data points representing the listeners listeners with largest and the smallest cohesion of judgements are connected by dotted lines. Solid line represents λ' index values calculated for the full set of data.

4. Conclusions

The present study has demonstrated the application of the λ' index and the KA' aggregate criterion (as well as their twin – λ and KA) for the analysis of judgements obtained from individual listeners or jurors. These measures may be especially helpful for ranking in cases of large variability of individual judgements. The examples of jury judgements during a musical competition as well as the judgements obtained in a psychoacoustical experiment both demonstrate that inter-subject variability of judgements may be high, even if the group of jurors or listeners is specially selected and seems to be homogenous.

Acknowledgment

The authors thank Dr. hab. Andrzej Miśkiewicz for providing raw experimental data used in Sec. 3.2 of this article.

References

[1] ARANOWSKA E., *Quantitative measurement in psychology* [in Polish], Scholar, Warszawa 2005.

[2] MANTURZEWSKA M., *On the difficult art of musical performances* [in Polish], [in:] *Judging musical performances*, AMFC, Warszawa 2006.

[3] MIŚKIEWICZ A., RAKOWSKI A., ROŚCISZEWSKA T., *Perceived roughness of two simultaneous pure tones*, Acta Acustica united with Acustica, **92**, 331–336 (2006).