Acoustic and Biomedical Engineering 2014

Perceptual Evaluation of the Effect of Threshold in Selective Mixing of Sounds

P. Kleczkowski* and M. Pluta
AGH University of Science and Technology, Department of Mechanics and Vibroacoustics
Al. A. Mickiewicza 30, 30-059 Krakow, Poland

Signal processing methods make possible such a mixing of signals that their overlapping in the time-frequency plane is reduced. This can be achieved by reducing the number of overlapping signals by discarding contributions from weaker signals and leaving only contributions from stronger ones. When applied to acoustic signals, this is referred to by the authors as selective mixing of sounds. Previous research has shown, that this rule, when applied to signals of musical instruments can provide some perceptual advantages over simple adding up the sound sources. In this paper, an experiment was carried out to determine the threshold of the value of relative energy of sound sources to control the decision about discarding a contribution from a particular sound source.

DOI: 10.12693/APhysPolA.125.A-117
PACS: 43.60.Hj, 43.60.+d, 43.75.Zz, 43.66.Lj

1. Introduction

Selective mixing of sounds is a specific technique developed for the production of musical recordings, which is currently under development. It is based on the reasoning related to the properties of hearing, briefly presented below.

The segregation of sounds from simultaneous sound sources becomes difficult for the ear when it is loaded with too much information. This paper is concerned with the following concept: given multiple acoustic sources, excessive information could be removed in those time-frequency regions, where numerous overlapping contributions from sources occur, by discarding some contributions, with the purpose that the remaining information from other contributions is segregated more effectively. This is performed by converting individual time signals representing sound sources into the time-frequency domain and then performing the comparison of energy of all signals in all time-frequency cells. In the utmost form of this processing, all contributions except the strongest one are discarded in each cell. This type of processing results in complete removal of spectro-temporal overlap between individual acoustic sources. Experiments have proved that the removal of large parts of signals of musical instruments or speech in the time-frequency domain may not be perceived in their mixture at all [1]. Other experiments in this area have revealed that the ear can use information contained in very small areas of the time-frequency plane, thus supporting the two known hypotheses in this field [2–6]. No minimum size of a time-frequency region that would contribute to perception of sounds was found.

While some attempts to use similar processing can be found in the literature, they were aimed at audibility of the effect [7, 8] or coding of audio signals [9, 10] but not at the improvement of quality of audio mixes.

The paradigm of selective mixing was presented in [11] and consequences of selective mixing in its utmost form of removal of spectral overlap in the case of speech signals were studied in [12]. Some listening experiments conducted by the authors indicated that statistically significant majority of listeners chose processed recordings as more detailed. A part of these experiments is reported in this paper.

For practical applications in audio engineering, a more useful version of selective mixing consists in removing a number of sound sources from a given time-frequency region, and leaving there several others, which contribute most to perception. A number of approaches for choosing sound sources to be discarded are possible. All of them lead to checking the value of energy of a sound source against a threshold, but the reference point of such threshold may be attained by different strategies. A straightforward one was assumed in this paper, and the effect of the value of a threshold on qualitative assessment of musical mixes was investigated by listening tests.

2. Formulation of the problem

A number of simultaneous sound sources can be arranged according to amplitudes of sounds they produce. The same can be done in any local region of the time-frequency plane. The principal mode of application of selective mixing is upon individual time-frequency cells. A hypothetical example of arrangement of contributions of energy from a number of sound sources in one cell is shown in Fig. 1.

The acoustic signals from independent sound sources are uncorrelated, therefore the total energy is equal to the arithmetic sum of individual energies:

$$S_T = \sum_{i=1}^{N} S_i. \tag{1}$$

The key issue in the application of selective mixing is

*corresponding author; e-mail: kleczkow@agh.edu.pl

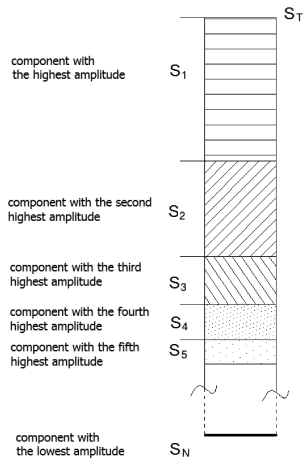

Fig. 1. An arrangement of simultaneous sound sources contributing to one time-frequency cell, according to their local amplitudes. S_i – value of energy of a sound source in the cell, S_T – total value of energy in the cell. Energy values are decreasing with increasing index, as suggested graphically.

how many sound sources should be discarded. Two basic approaches are possible.

1. In each time-frequency cell there is the same number of contributing sound sources. That number is fixed for the entire audio recording, for example two, three or more. The sources chosen are taken successively from the ordered list (as in Fig. 1), starting from the one with the highest amplitude.

2. In each time-frequency cell there is a specific number of contributing sound sources.

Option 2 provides more flexibility. As only an integer number of sources can be used, differences between two, three or four sources are meaningful and thus adjustments in approach 1 can only be applied in coarse steps. Therefore, the decision must be taken for each individual cell.

There is a number of possible approaches to this decision, but a straightforward one is based on the ratio of energy of each particular source S_i to the sum of energies of all sources S_T

$$t = \frac{S_i}{S_T}. \tag{2}$$

For the ease of conventional use in acoustics this can be expressed in decibels:

$$r\ [\mathrm{dB}] = 10 \log \frac{S_i}{S_T} = 10 \log S_i - 10 \log S_T. \tag{3}$$

In the practical computational procedure, in each cell those sources are discarded energies of which are below a threshold r.

As the perceptual results of selective mixing are highly subjective, there exists no method to determine the appropriate value of r in an analytic way. Presumably, this will depend on the musical material to be mixed. In this paper, the value of r was evaluated for two similar pieces of music, by experimental assessment carried out by a panel of listeners.

3. Stimuli

Two sets of sound tracks were prepared on the basis of excerpts from two different pieces of instrumental jazz music, further referred to as "jazz1" and "jazz2". Both sets were mixes of eight tracks: two keyboards, guitar, bass guitar, saxophone, kick drum, snare drum and overhead. "Jazz1" lasted 8 seconds. "Jazz2" was a different piece of music played by the same musicians on the same instruments and lasted 10 seconds. The sets, i.e. mixes, differed only in the mixing method used. Each mix was prepared in four versions: one conventional mix ("original") and three different variants of selective mix. Parameter r (threshold) in selective mixes was set to three different values: –6 dB, –8 dB, and –12 dB.

In order that perceptual effect is favourably assessed by listeners, the spectrograms of original sounds (tracks) need to be smoothed [13]. The detailed parameters of smoothing have little influence on perception, therefore just perceptually acceptable parameters were used in preparation of stimuli for this paper.

The process of selection reduces the total energy of audio mixes, as some sound sources are discarded. Therefore, RMS of all generated samples within a set was normalised to avoid any bias resulting from uneven loudness of mixes.

4. Procedure

Participants of the test were asked to listen to and evaluate selected aspects of a number of sound samples. The test was performed in silenced rooms of a recording studio. Each participant worked separately, using a dedicated computer software that controlled the procedure. Computers were equipped with external audio interfaces: M-Audio Fast Track Pro or M-Audio Fast Track Ultra 8R, and stimuli were reproduced by means of closed headphones Beyerdynamic DT 770 Pro. The original tracks and mixes were stereophonic.

The test was attended by two groups of participants. The first one (Group I) consisted of 33 students of the major of acoustical engineering. This group evaluated "jazz1". The second one (Group II), consisting of 20 students of the same speciality, evaluated "jazz2". The groups did not intersect. Group I has undergone audiogram examination, and only one listener mildly exceeded the 20 dB limit at two frequencies, but his results were not excluded. Audiograms of the participants of Group II were not obtained, but none of them reported any problems in hearing. The test procedure was similar in case of both groups.

The function of the software was to present, upon a listener's request, a chosen sound sample, and record the answers. Depending on the group, all presented samples were either "jazz1" or "jazz2". Samples were named with letters from "A" to "D" denoting the type of processing: original (unprocessed), –6 dB, –8 dB, and –12 dB, but the assignment was random for each of the two groups.

Users were not informed about the nature of differences between samples. Each subject was asked to listen to each sample at least 16 (the first group) or 20 (the second group) times in any order and evaluate each sample using an integer scale from 0 (very poor) to 5 (excellent) in each of six categories: spaciousness, localisation of sources, clarity, lack of distortion, lack of noise, and general impression.

The test could have been completed only if each sample was evaluated in every category. By that time the listener was allowed to change his/her evaluation any number of times.

Due to the nature of the procedure, no training session was necessary. There was no time limit and no upper limit for the number of replays. Test sessions lasted about 15 minutes in the first and 25 minutes in the second group.

5. Test results

Due to the fact that two different sets of samples were used during the test, the results were analysed separately in both groups (Table). In the graphs (Figs. 2–8) Series I and Series II represent results obtained with Group I and Group II, respectively.

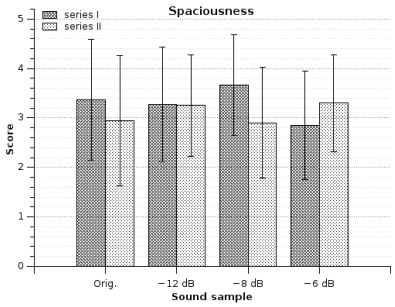

Fig. 2. Average scores and standard deviations of the sound quality evaluation in the category "spaciousness", depending on the value of threshold r in both groups of listeners.

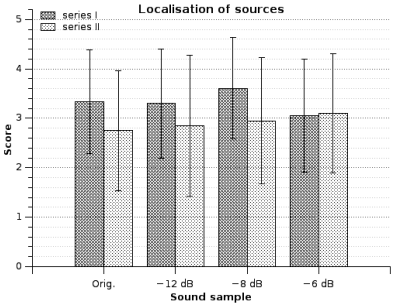

Fig. 3. Average scores and standard deviations of the sound quality evaluation in the category "localisation of sources", depending on the value of threshold r in both groups of listeners.

Generally, results in Group II display lower scores, except for some categories in case of –6 dB threshold (spaciousness, localisation of sources, and general impression)

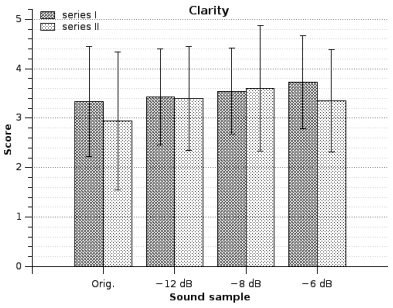

Fig. 4. Average scores and standard deviations of the sound quality evaluation in the category "clarity", depending on the value of threshold r in both groups of listeners.

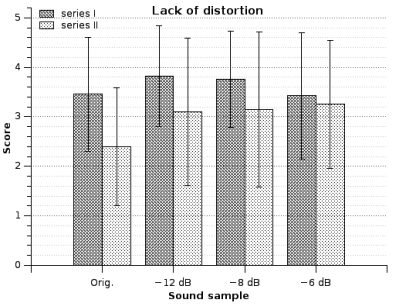

Fig. 5. Average scores and standard deviations of the sound quality evaluation in the category "lack of distortion", depending on the value of threshold r in both groups of listeners.

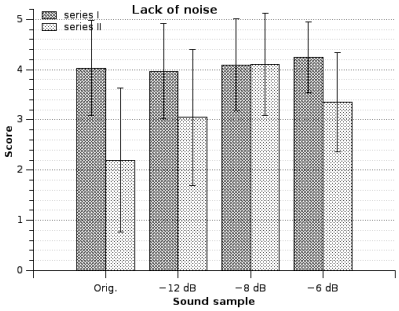

Fig. 6. Average scores and standard deviations of the sound quality evaluation in the category "lack of noise", depending on the value of threshold r in both groups of listeners.

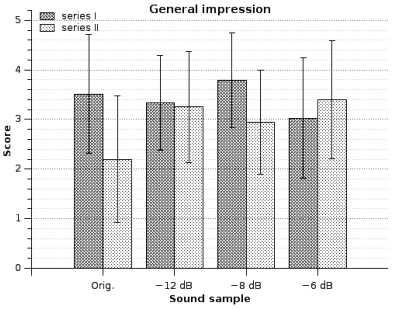

Fig. 7. Average scores and standard deviations of the sound quality evaluation in the category "general impression", depending on the value of threshold r in both groups of listeners.

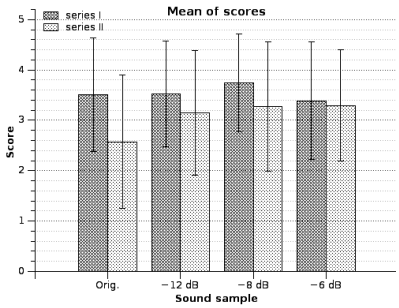

Fig. 8. The mean of scores, depending on the value of threshold r in both groups of listeners.

and some in case of –8 dB (clarity, lack of noise). In Group I, the highest mean score was assigned to the variant with –8 dB threshold, but the differences in the mean scores among categories were statistically insignificant, except the difference between the scores for –8 dB (highest score) and –6 dB (lowest score) which is significant at $p < 0.02$ (normal distribution assumed). In Group II, the highest mean score was assigned to the version with –6 dB threshold, with nearly the same score for –8 dB version and slightly lower score for –12 dB version. The differences between these versions were statistically insignificant, while the difference between each of them and the original version (conventional mix) was significant. For differences between –6 dB and –8 dB threshold versions and the original version the significance level was $p < 0.01$, compared to $p < 0.05$ for differences between the –12 dB and the original version (t distribution).

TABLE

Average scores and standard deviations of the sound quality evaluation in six categories, depending on the value of threshold r.

r [dB]	Spaciousness	Localisation of sources	Clarity	Lack of distortion	Lack of noise	General impression	Mean score
Group I							
$-\infty$ (unprocessed)	3.36 ± 1.22	3.33 ± 1.05	3.33 ± 1.11	3.45 ± 1.15	4.03 ± 0.95	3.52 ± 1.20	3.51 ± 1.13
–12	3.27 ± 1.15	3.30 ± 1.10	3.42 ± 0.97	3.82 ± 1.01	3.97 ± 0.95	3.33 ± 0.96	3.52 ± 1.05
–8	3.67 ± 1.02	3.61 ± 1.03	3.55 ± 0.87	3.76 ± 0.97	4.09 ± 0.91	3.79 ± 0.96	3.74 ± 0.97
–6	2.85 ± 1.09	3.06 ± 1.14	3.73 ± 0.94	3.42 ± 1.28	4.24 ± 0.71	3.03 ± 1.21	3.39 ± 1.17
Group II							
$-\infty$ (unprocessed)	2.95 ± 1.32	2.75 ± 1.21	2.95 ± 1.39	2.40 ± 1.19	2.20 ± 1.44	2.20 ± 1.28	2.58 ± 1.32
–12	3.25 ± 1.02	2.85 ± 1.42	3.40 ± 1.05	3.10 ± 1.48	3.05 ± 1.36	3.25 ± 1.12	3.15 ± 1.24
–8	2.90 ± 1.12	2.95 ± 1.28	3.60 ± 1.27	3.15 ± 1.57	4.10 ± 1.02	2.95 ± 1.05	3.28 ± 1.28
–6	3.30 ± 0.98	3.10 ± 1.21	3.35 ± 1.04	3.25 ± 1.29	3.35 ± 0.99	3.40 ± 1.19	3.29 ± 1.10

If results of both groups are combined (Figs. 9–10), the best score of 3.566 is obtained for the –8 dB version, and the second best for –12 dB version. The lowest joint score is for the original: 3.154. The difference between the means for –8 dB version and the original is significant at $p < 0.01$ (normal distribution). The results shown in Figs. 9 and 10 were obtained by averaging weighted with numbers of listeners in both series.

The results in particular categories are less consistent.

Susceptibility of individual categories to the value of threshold r was examined by calculating the variance of results in each of the categories. Each term in variance sum was calculated as weighted average of the results in series I and II. The susceptibility thus obtained is shown in Fig. 11.

Distinctively highest variance was obtained in the category "lack of noise". In this category the weighted mean score (4.094) was highest for $r = -8$ dB. The lowest score was assigned to unprocessed version, and the difference between the two was statistically significant. This difference was particularly pronounced in series II. An

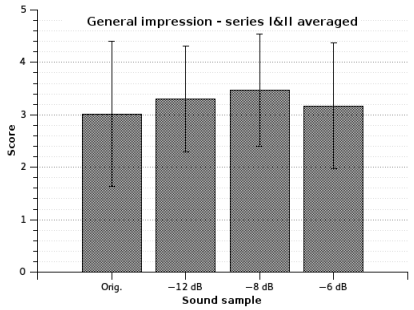

Fig. 9. General impression evaluation depending on the value of threshold r averaged in both groups of listeners.

unexpected interpretation of this result is that the perception of the original versions were perceived as noisy, when compared to selective versions. The second highest variance was for lack of distortion, and comparison with data in Table I indicates that the processed versions were perceived as less distorted. The lowest value of variance was obtained for localisation.

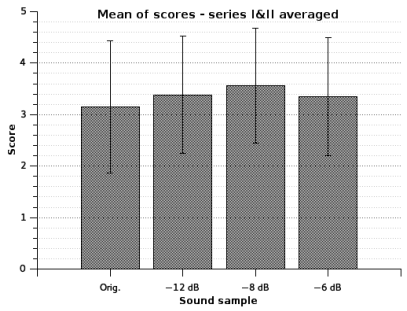

Fig. 10. The mean of scores depending on the value of threshold r averaged in both groups of listeners.

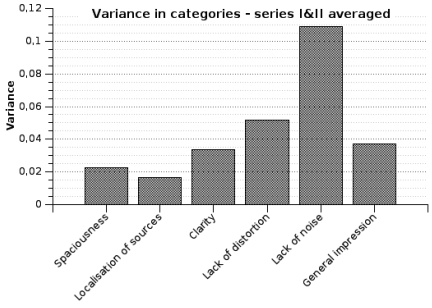

Fig. 11. The variance of results in individual categories. Terms in variances were weighted averages of the results in series I and II.

6. Conclusions

The combined results for series I and II, as far as the mean of all scores and the most important category of general impression are concerned, indicate the value of threshold r of –8 dB as the most favourable. From these results it can be concluded, that for the musical mixes with a similar number of simultaneous sound sources, the level of threshold r of –8 dB seems to be close to optimal one. There were eight tracks in the investigated stimuli but as is the case in most kinds of music, not all sound sources sounded at individual instants of time and the number of currently active sources was varying.

The second most preferred value of the threshold when averaged over series I and II was -12 dB. There were some differences between the series. In series I, the most preferred value of the threshold was –8 dB, while in series II it was –6 dB.

The differences between preferences for particular thresholds were not considerable. In series I, only the difference between –8 dB and –6 dB was statistically significant, while in series II all processed versions (–6, –8 and –12 dB) were found significantly better than the unprocessed version (corresponding to threshold $r = -\infty$). The advantage of the –8 dB version versus the unprocessed version is also significant when the results of series I and II are averaged.

The perception of the effect was demonstrated to be sensitive to the value of threshold, as its change from –8 dB to –6 dB resulted in the shift of the mean score from the highest to the lowest in series I.

The comparison of variances in particular categories of assessment has demonstrated that the category "lack of noise" was the most susceptible to the value of threshold r. The difference in ratings between $r = -\infty$ and $r = -8$ dB in that category was statistically significant and indicated, that listeners perceived selectively mixed versions as less noisy.

References

[1] P. Kleczkowski, *Arch. Acust.* **37**, 355 (2012).

[2] N.F. Viemeister, G.H. Wakefield, *J. Acoust. Soc. Am.* **90**, 858 (1991).

[3] P.A. Howard–Jones, S. Rosen, *Acustica* **78**, 258 (1993).

[4] P.A. Howard–Jones, S. Rosen, *J. Acoust. Soc. Am.* **93**, 2915 (1993).

[5] M.P. Cooke, *J. Acoust. Soc. Am.* **119**, 1562 (2006).

[6] J. Barker, M. P. Cooke, *Speech Comm.* **49**, 402 (2007).

[7] M.C. Kelly, A.I. Tew, *The Continuity Illusion in Virtual Auditory Space*. Proc. 113th Conv. Audio Eng. Soc., Preprint 5548 (2002).

[8] M.C. Kelly, A.I. Tew, *The Significance of Spectral Overlap in Multiple-source Localization*. Proc. 114th Conv. Audio Eng. Soc., Preprint 5725 (2003).

[9] C. Faller, F. Baumgarte, *Binaural Cue Coding Applied to Stereo and Multi-Channel Audio Compression*, 112th Conv. Audio Eng. Soc., Preprint 5574 (2002).

[10] C. Faller, F. Baumgarte, *Binaural Cue Coding Applied to Audio Compression with Flexible Rendering*, 113th Conv. Audio Eng. Soc., Preprint 5686, (2002).

[11] P. Kleczkowski, *Selective Mixing of Sounds*, 119th Conv. Audio Eng. Soc., Preprint 6552 (2005).

[12] P. Kleczkowski, M. Plewa, M. Pluta, *Increasing Intelligibility of Multiple Talkers by Selective Mixing*, 129th Conv. Audio Eng. Soc., Preprint 8309 (2010).

[13] P. Kleczkowski, A. Kleczkowski, *Advanced Methods for Shaping Time-Frequency Areas for the Selective Mixing of Sounds*, 120th Conv. Audio Eng. Soc., Preprint 6718 (2006).