PAN ARCHIVES OF ACOUSTICS
Vol. **37**, No. 3, pp. 355–363 (2012)

DOI: 10.2478/v10168-012-0045-0

Perception of Mixture of Musical Instruments with Spectral Overlap Removed

Piotr KLECZKOWSKI

AGH University of Science and Technology
al. A. Mickiewicza 30, 30-059 Kraków, Poland; e-mail: kleczkow@agh.edu.pl

(*received June 27, 2011; accepted July 24, 2012*)

The issue of auditory segregation of simultaneous sound sources has been addressed in speech research but was given less attention in musical acoustics. In perception of concurrent speech, or speech with noise, the operation of time-frequency masking was often used as a research tool. In this work, an extension of time-frequency masking, leading to the removal of spectro-temporal overlap between sound sources, was applied to musical instruments playing together. The perception of the original mixture was compared with the perception of the same mixture with all spectral overlap electronically removed. Experiments differed in the method of listening (headphones or a loudspeaker), sets of instruments mixed, and populations of participants. The main findings were: (i) in one of the experimental conditions the removal of spectro-temporal overlap was imperceptible, (ii) perception of the effect increased when removal of spectro-temporal overlap was performed in larger time-frequency regions rather than in small ones, (iii) perception of the effect decreased in loudspeaker listening. The results support both the multiple looks hypothesis and the "glimpsing" hypothesis known from speech perception.

Keywords: sound segregation, spectral overlap, spectrogram, auditory scene analysis, time-frequency mask, multiple looks, glimpses.

1. Introduction

The auditory system's mechanisms for extracting sounds from separate sources rely on spatial, time and spectral parameters. It is well known that the segregation task becomes more difficult when spectro-temporal patterns of sounds overlap. Investigation on how the ear copes with overlapping sounds is a difficult task, as it involves: non-linear addition of masking, a sort of central masking called informational masking, illusion of continuity, binaural hearing and mechanisms collectively referred to as Auditory Scene Analysis (ASA) (Bregman, 1990). The segregation of overlapping sounds, besides ASA, has been investigated within research on speech perception. In ASA an analysis-synthesis process is assumed, where the acoustic scene is first decomposed into a set of segments, which are then grouped to form coherent and independent streams in a synthesis process.

Attempts have been made towards segregation of sounds by computational means. They are collectively referred to as Computational Auditory Stream Analysis (CASA, for a review, see Wang and Brown, 2006). An often used computational paradigm in CASA is to estimate time-frequency (t-f) mask, where "mask" denotes the operation of applying a mask. Masks are applied to spectrograms of mixed sounds. If the value of 1 is applied for a t-f unit in which the target energy is stronger than the total interference energy, and the value of 0 otherwise, the mask is called ideal binary mask (Wang, Brown, 2006; Brungart *et al.*, 2009).

Auditory segregation of overlapping sounds when the sources are musical instruments was given less attention, and concentrated on analysing sequences of pitches (Bregman, 1990). Kelly and Tew (2002, 2003) applied the operation of masking with coefficients varying from 0 to 1 to two musical instruments in a binaural recording. Their main finding was that it was possible to remove the weaker signal in a particular t-f location only if its level was lower by at least 15 dB, and no perceptible degradation of the signal should occur. In research on speech perception, including the applications of CASA, the speaker is a target and the background (typically noise) is a masker. In the case of the cocktail party effect one voice is a target and the others are a collective masker. Investigation on

segregation of musical instruments requires a different approach: contributions of individual instruments (or at least their groups) are all targets. Another problem in approaching segregation in musical acoustics is that it is difficult to construct stimuli so that perception is quantifiable.

This author has developed an appropriate modification of binary t-f masking, where all spectro-temporal overlapping between all individual sound sources is removed. This processing algorithm removes most of the energetic masking between individual sound sources. That property of signals changes conditions of segregation for the ear so it may have perceptual implications and is worth investigation. It has been applied both to musical instruments playing together (Kleczkowski, 2005, 2008) and to monaural mixes of multiple talkers (Kleczkowski, Pluta, 2012). The earlier works reported qualitative perceptual properties of this operation, while the latter showed that it did not change the rate of understanding of concurrent speech. The aim of this paper is to find basic quantifiable properties of this operation when performed on musical instruments playing together.

The removal of spectro-temporal overlap (RSO) is substantially different from removal of elements of sounds resulting from algorithms of lossy compression of audio signals. The RSO algorithm operates on separate sound sources, while compression techniques operate on the mixed signal. It can also be easily shown, that considerably more energy is removed from sounds with RSO than with lossy compression.

There were four objectives of this work: (i) to measure the perceptual difference between the natural and the spectrally non-overlapping presentations of musical instruments playing together, (ii) to evaluate the effect of the average size of t-f regions where overlapping is eliminated on this difference, (iii) to evaluate the relation between the perceived difference and the ratio of retained sound to removed sound, in terms of both energy and t-f surface, and (iv) to investigate whether the multiple looks and glimpses hypotheses also hold for perception of musical instruments and if so, to obtain any assessment of the size of "glimpses". The multiple-looks hypothesis (Viemeister, Wakefield, 1991) postulates that the ear is capable of integrating auditory percepts from small elements scattered in time. This hypothesis was confirmed within research on speech perception, and extended to the time-frequency domain, where the ear is supposed to analyse scattered spectro-temporal components of sounds (referred to as "glimpses") to perform segregation (Howard–Jones, Rosen, 1993a, 1993b; Cooke, 2006; Barker, Cooke, 2007; Lu, Cooke, 2008).

In experiment 1, listeners compared original mixes of musical instruments with their RSO versions in a psychophysical experiment over the headphones. The aim of experiment 2 was to test how the stimuli of experiment 1 were perceived in loudspeaker presentation, i.e. in settings typically encountered when listening to music.

2. Processing and stimuli

2.1. Implementation of the removal of spectro-temporal overlap

In the rest of this paper, a single and smallest possible element of a t-f distribution will be referred to as a "cell", while a group of neighbouring cells is referred to as a "region".

To remove spectro-temporal overlap between sound sources their separate acoustic signals are needed. After the t-f distributions (spectrograms) of all input signals are obtained, the spectrograms are compared cell by cell, and in each cell the signal characterised by the highest value of amplitude is chosen

$$|F|_{k,n,\mathrm{out}} = \max\left\{|F|_{k,n,1}, |F|_{k,n,2}, \ldots |F|_{k,n,p}\right\}, \quad (1)$$

where F is a t-f coefficient, k is the index of a frequency bin, n is the index of a time frame, p is the number of acoustic sources, and "out" denotes an output t-f signal. The argument of the modulus function on the left is passed to the output t-f signal. The operation in (1) was performed by a simple algorithm for finding the maximum element in a set. Cells belonging to other sounds in that t-f location were not passed to the output signal.

The operation in (1) is a "winner takes all" competition that takes place in each t-f cell, i.e. between contributions from all sound sources. An example of the occupancy map in the t-f plane resulting from RSO processing of two musical signals is presented in Fig. 1. As can be seen in this figure, the RSO results in a partition of the spectrogram of the mixed signal: some cells or regions contain only the contribution of one instrument, while the others only the contribution of the other one. There is no cell containing both contributions, thus there is no t-f overlap of sound sources.

The t-f analysis/synthesis method used was based on the Modified Discrete Cosine Transform (Malvar, 1992; Kleczkowski, 2002) – a perfectly invertible block t-f transform. A custom software for the entire procedure was written by the author in C.

Due to the properties of block transforms, the sizes of t-f cells had to follow a fixed grid. The duration of the individual t-f cell and its related bandwidth was chosen at 11.6 ms/43.06 Hz, with reasonable alternatives of 5.8 ms/86.12 Hz or 23.2 ms/21.53 Hz. The chosen size was a compromise between the duration of the auditory "time window", estimated by Moore *et al.* (1988) at around 8.3 ms at 500 Hz and 8 ms at 2000 Hz, and the frequency width of a t-f cell. The latter should be substantially narrower than a local

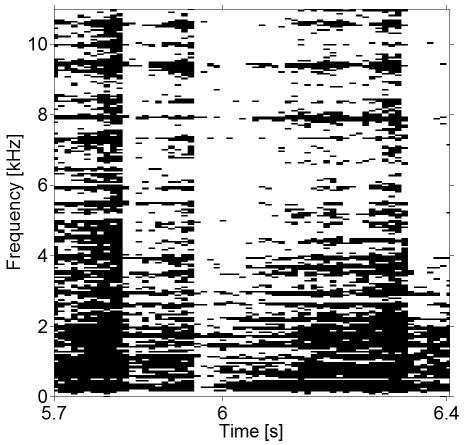

Fig. 1. The map of occupation of t-f cells after performing RSO operation according to (1) on spectrograms of two simultaneously playing instruments: drums and a synthesizer ("dr-sn" set in Table 1) . Black indicates all t-f cells where the synthesizer "won the competition", i.e. its amplitude was bigger that that of the drums. In the re-synthesized RSO mix, all t-f cells marked black in this figure contain only the contribution of the synthesizer. White indicates the opposite situation (amplitude of drums was bigger of the two). The map shows a 1.7 s long excerpt used in experiment 1. The frequency axis is limited to 11 kHz in order to increase the vertical resolution of the plot.

critical band (CB), even for lowest frequencies. With the first alternative, the bandwidth in low frequencies (86.12 Hz) would be close to one CB.

As can be seen in Fig. 1, a considerable part of the t-f occupancy map is covered by small scattered t-f regions, often consisting of a single t-f cell. By smoothing operations the average size of the t-f region can be made bigger. One of the objectives of this work (ii) was to investigate the relation between this size and perception of RSO. Therefore two modes of processing were used. The first will be referred to as individual cells mode: t-f regions are identical to t-f cells. The other mode, referred to as clustered cells mode, involved clustering of individual t-f cells into larger regions. Clustering was performed on individual spectrograms, prior to operation (1). It was based on local concentration of t-f energy. Clusters were formed as the result of two-dimensional averaging over the dimensions of time and frequency. Due to stochastic nature of this process clusters had no fixed dimensions nor did they always form compact shapes in the plane, but the number of small scattered regions or individual cells was considerably reduced. After clustering the RSO algorithm worked the same way as in individual cells mode using (1), just rendering occupancy maps that were smoother. More details on clustering are given in (Kleczkowski, 2008). Figure 1 shows an example of RSO mix in individual cells mode, while Fig. 2 shows the same mix in clustered cells mode. Two modes of RSO processing presented above also serve objective (iv) of this paper, i.e. the assessment of the size of "glimpses" in time and frequency.

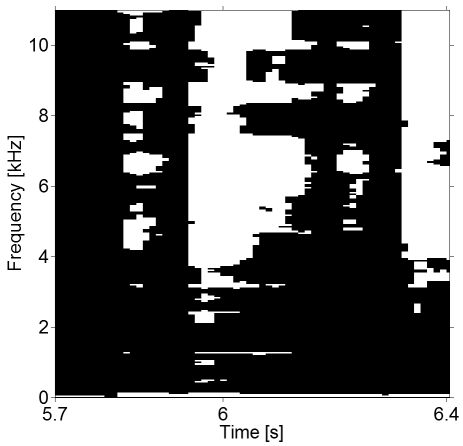

Fig. 2. The map of occupation of t-f regions by the same instruments as in Fig. 1 (the synthesizer – black, drums – white), after RSO in clustered cells mode.

2.2. Stimuli

The sound sources chosen for experiments were professional recordings of musical instruments: bass, two guitars (recorded in one track), drums, synthesizer and saxophone playing the same fragment of a pop-jazz piece. An excerpt lasting 7 s was chosen. Monophonic tracks (16 bits, 44.1 kHz) containing individual instruments were mixed in 12 combinations. The bass was included only in the mix of all five instruments, as it occupied the low end of the spectrum, with little spectral overlap with other instruments. Mixes included all possible combinations of two and three instruments, bass being excluded. Prior to mixing, the relative levels of all instruments were adjusted by a professional audio engineer, so that an appropriate balance of instruments was obtained. Figure 3 presents the t-f oc-

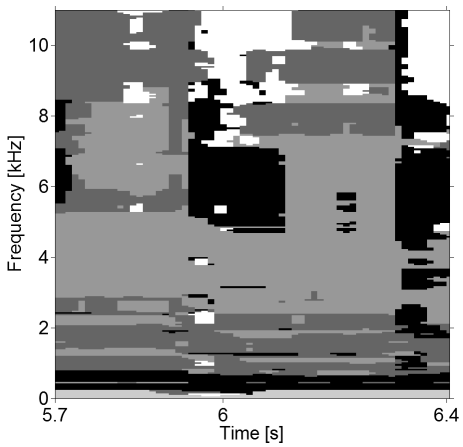

Fig. 3. The map of occupation of t-f regions by the sounds of drums – white, bass – light grey, the saxophone – grey, the synthesizer – dark grey, and guitars – black, i.e. the complete band ("all_5" set), after RSO in clustered cells mode.

cupancy map of the RSO mix of all five instruments, further referred to as "all_5", used in the experiment in the clustered cells mode.

For each of 12 sets of instruments, three stimuli were prepared: (I) the original mix, obtained by converting the original tracks to the t-f domain and mixing them in that domain to control for possible effects introduced by t-f processing, (II) RSO mix in the individual cells mode, (III) RSO mix in the clustered cells mode. Since RSO processing reduces the energy of all sound sources, stimuli (II) and (III) were normalised to have an RMS value equal to that of (I). The stimuli were also normalized across all sets.

Table 1 lists all stimuli used with their percentage of energy and t-f area retained after RSO. All stimuli were monaural. Examples of maps of occupation of the t-f plane are presented in Figs. 1 to 3. These figures show a 1.7 s long excerpt used in experiment 1. The frequency axis is limited to 11 kHz in order to increase the vertical resolution of the plot.

Table 1. Sets of instruments used as stimuli and their respective percentages of energy and efficient area retained in the mix after RSO. Energy and area values are averaged over instruments in the set. Symbols: b – bass, g – guitars, dr – drums, sn – synthesizer, sx – saxophone, IC – individual cells, CC – clustered cells.

Set of instruments	% energy retained		% area retained	
	IC	CC	IC	CC
dr-g	86.9	78.4	63.8	70.3
g-sx	87.7	82.5	70.5	64.2
dr-sx	88.8	83.6	60.3	65.5
dr-g-sx	79.3	70.5	46.1	49.7
sn-sx	83.8	79.1	65.9	61.2
dr-sn	85.8	82.2	62.9	67.9
dr-g-sn-sx	67.3	58.8	36.8	38.6
dr-sn-sx	76.7	70.5	45.1	46.8
dr-g-sn	71.5	63.9	46.8	50.3
g-sn	76.1	70.7	66.7	61.7
all_5	68.3	61.7	31.9	33.5
g-sn-sx	71.2	64.3	51.9	45.5

The percentage of energy retained was computed separately for each of the instruments in each of the stimuli. The values presented are means over all instruments in a given stimulus. The percentages of area retained were computed according to a similar rule. In the calculation of energy retained, all coefficients of the t-f distribution of a given instrument contributed to the calculation of the proportion's denominator. This method was considered improper for the calculation of area, since a considerable number of t-f coefficients of each instrument contained only background noise. Their effect was negligible in the calculation of energy, but would introduce a bias in the calculation of area. Therefore a threshold was set for a t-f coefficient to be included in the denominator: $|F| \geq 0.0003 \cdot \max\{|F|\}$. This threshold corresponded to -70 dB relative to the coefficient with the highest value, an approximation of the signal-to-noise ratio in the recording process. The coefficients thus selected contained about 99% of the energy of an instrument, and the corresponding area will be referred to as the effective area. The percentage of t-f area computed in this study is similar to the "visibility" parameter used by Barker and Cooke (2007), except that in the latter no correction for background noise was included.

It can be noticed in Table 1, that the least area retained is for the "all_5" set, and "dr-g-sn-sx" (four instruments) comes next. This can be easily explained: the effect of RSO is that the t-f plane has to be divided between competing sound sources. The more sources, the less average area remains available for each of them. This is less noticeable for energy, as each instrument has its own energy.

All stimuli were generated offline on a PC. They were stored as audio files at 16 bits, 44.1 kHz resolution.

3. Experiment 1

3.1. Choice of stimuli

A pilot test for this experiment participated by expert listeners showed that the perceptual difference between original mixes and RSO mixes increased substantially in the clustered cells mode. It also tended to increase with the number of instruments mixed i.e. with the reduction of energy and t-f area retained in RSO mixes (cf. Table 1, that has been ordered according to results of the pilot test). Two sets were chosen for experiment 1: "all_5" for its lowest values of energy and area retained in RSO mix, and "sn-sx", as its overall RSO detectability was medium while the discrepancy between its results for the individual and clustered cells modes was high. Out of the original 7 s long excerpt, a shorter 1.7 s long excerpt was used in experiment 1, as a compromise between eliminating memory-related aspects of the experiment and the length required to assess a musical material.

3.2. Subjects

Nineteen subjects aged 20-22, all of them students of the Acoustic Engineering course at the AGH University of Science and Technology, participated in this experiment. All listeners had normal hearing defined as thresholds within 20 dB of nominal at octave frequencies from 250 through 8000 Hz. The thresholds were measured by Békésy audiometry using headphones. All had at least some experience in psychoacoustics tests.

3.3. Experimental setup

Stimuli were played out from a PC netbook through a M-Audio Fast Track Pro USB audio interface. They were presented to the subject through Beyerdynamic DT 770 Pro closed headphones. Stimuli were monophonic but the presentation was diotic, as the listeners found monaural presentation tiring. An attempt to set a common level failed as some listeners found it too low or too high. The presentation level was set individually to a level that was comfortable for each listener during a practice session. The starting level was 75 dB SPL and the adjustment range allowed did not exceed ±6 dB. The listeners were seated in a sound isolated room fitted with sound absorbing material. The screen of a notebook PC contained control buttons to play/stop stimuli and to give answers. The software was developed by dr Marek Pluta of AGH University.

3.4. Experimental procedure

The same–different task was used, in 1AFC (One Alternative, Forced Choice) mode (Kingdom, Prins, 2010). Each trial consisted of a pair of stimuli: the original mix and the RSO mix, in random order, or two identical stimuli. They were separated by a pause of 200 ms. The subjects' task was to press one of two keys on a PC screen: "same" or "different". Feedback was given after each trial, i.e. the subject was informed in the computer screen whether his response was correct or false. This is recommended in experiments measuring sensitivity. The subject activated a next trial by pressing the "next" key. Tests for both sets: "all_5" and "sn-sx" were held separately, with a short break. Each of two tests consisted of 240 trials. 120 trials contained pairs of identical stimuli (original mixes and RSO mixes in both modes). The other trials contained an original mix and an RSO mix, with half of these pairs containing the individual cells mode and the other half containing the clustered cells mode. The sequence of different stimuli within the entire lot of 240 was random, but was held fixed for all subjects. Prior to the main experiment, each subject took a practice session of 30 trials.

3.5. Results and discussion

For each of the four experimental conditions and for each subject, the index of stimulus detectability d' was calculated according to (Gescheider, 1997):

$$d' = z(H) - z(F), \tag{2}$$

where H is hit rate i.e. the rate of detection of differences, and F is false alarms rate i.e. the rate of identical stimuli incorrectly classified as different, z denotes H or F rate converted to the location along the abscissa of standarized normal distributions, where $z(F)$ is a location along the noise distribution and $z(H)$ along the signal-plus-noise distribution. The main application of the d' index is to compare detectabilities of different stimuli, but it is usually assumed that $d' = 1$ is a threshold value, as this value corresponds to 76% correct recognitions in 2AFC (Two Alternatives, Forced Choice) tasks (Moore, 2003). Hence, a value below 1 indicates that the stimulus was not detectable.

Table 2 presents mean d' values for investigated stimuli, averaged over all subjects. The results in four experimental conditions can be summarised as follows. In individual cells mode the RSO effect was imperceptible with "sn-sx" stimulus (this condition will be further referred to as "imperceptible condition"). In the same mode with "all_5" stimulus perception was close to the threshold. In clustered cells mode the effect was perceptible with "sn-sx" stimulus and easily perceptible with "all_5"stimulus. There was a considerable spread in the results among listeners.

Table 2. Results of experiment 1: mean values of detectability index d' for investigated stimuli, averaged over all subjects; IC – individual cells mode, CC – clustered cells mode, σ – standard deviation.

Set of instruments	IC		CC	
	mean d'	σ	mean d'	σ
sn-sx	0.46	0.43	1.35	0.72
all_5	0.95	0.65	2.46	1.19

The comparison of data in Table 1 with data in Table 2 indicates that the decrease of energies retained in sounds after RSO processing increases the rate of recognitions of such processed mixes. This is the case in all four possible paired comparisons: the sn-sx set versus the all_5 set in both modes, and the individual cells mode versus the clustered cells mode for both mixes. When t-f areas are compared, three out of four comparisons support the statement that the decrease of t-f area retained increases the rate of recognition. The only exception is the all_5 set, where the increase in the rate of recognition is associated with a relatively slight increase of t-f area. This can be accounted for by random factors in the proportions of areas. The rule observed can be simply accounted for: the removal of energy and t-f area distorts sounds.

The condition of individual cells mode applied to "sn-sx" set, where RSO is imperceptible, can be used to demonstrate that the ear is able to perform the fusion of the auditory scene from the mixture of sounds altered by RSO processing. The term "fusion" is used here in the meaning of building a consistent percept out of some distinct parts.

It is doubtful that the continuity illusion is the basis for fusion, as the interrupting sound must exceed the interrupted sound at least by an amount causing complete masking (Bregman, 1990). In RSO, all sound components are removed when they fall just below 0 dB related to the masker. However, continuity illusion can support fusion.

Informational Masking (IM) seems to have an important contribution. The hypothetical role of IM can be the following. If RSO removes some spectro-temporal parts of sounds which would not be masked energetically by other sound sources in the mixture, then the difference brought by RSO should be perceptible. This removal takes place in the imperceptible condition. Therefore there must be some other factor or factors that make RSO in this condition imperceptible. The supposition that it is IM is supported by a known property of informational masking: it is stronger when both the target and masker are presented to the same ear. It is also supported by investigation on informational masking within a sound of one instrument (Kleczkowski *et al.*, 2010).

Another important conclusion from the existence of an imperceptible condition is that the multiple looks/glimpses hypotheses can hold for perception of musical instruments. However, as can be seen in Table 2, the ear is not completely successful in applying glimpses. In the remaining paragraphs of this section an attempt is made towards quantitative analysis of the frequency widths of "glimpses" in the imperceptible condition ("sn-sx", individual cells).

The number of contiguous cells (i.e. belonging to one sound source) along the frequency axis was assumed the width of a "glimpse". The t-f decomposition used the linear frequency scale, while a measure of bandwidth should use a perceptually justified frequency scale. The critical band was used as a unit of measure of widths of "glimpses". Thus, each counted number of contiguous cells n was converted to linear frequency bandwidth: $n \cdot 43.06$ [Hz], and that bandwidth was converted to a fraction of a local CB: $n \cdot 43.06$ /CB_l [Hz], where CB_l denotes frequency width of a local CB, according to the table by Zwicker (Fastl, Zwicker, 2007).

In order to present the results as a histogram, all resulting values were grouped into eight ranges: from 0 to 15% of CB_l, from 15% to 30% of CB_l, ..., from 90% to 105% of CB_l. All values greater than 105% of CB_l were included in one common group. The histogram of the results is shown in Fig. 4. More details on computation are included in the Appendix.

In individual cells mode 64% of values fell into the 0–15% of CB_l range, in clustered cells mode that proportion was 16%. The median of the results in individual cells mode was 10.2% of CB_l and the median in clustered cells mode was 43.9% of CB_l.

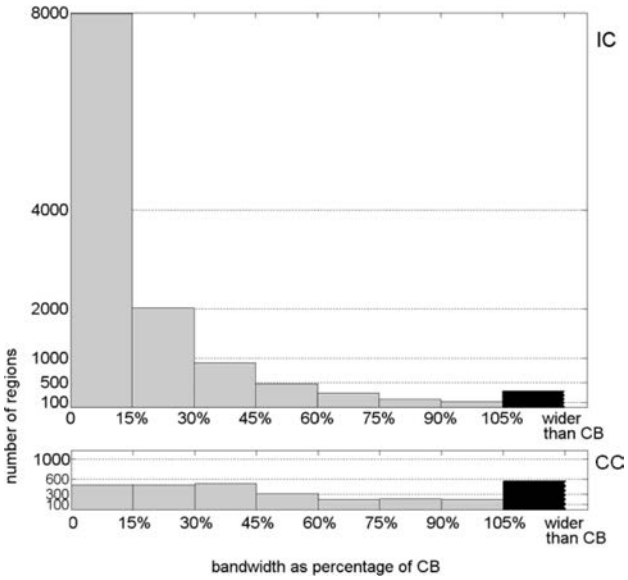

Fig. 4. Upper: the histogram of numbers of t-f regions in a particular range of frequency widths expressed as a percentage of a local CB, in individual cells mode. Lower: the analogous histogram in clustered cells mode. IC – individual cells mode, CC – clustered cells mode.

These results indicate that the hypothesis of "glimpses" holds better with narrow frequency widths of "glimpses". It has been demonstrated that with this particular stimulus (i.e. in the imperceptible condition), when 64% of "glimpses" have widths below 15% of CB_l the multiple looks/glimpses hypothesis holds perfectly. However, this statement must be accompanied by other conditions of experiment 1: 83.8% of the original energy and 65.9% of original t-f area were preserved in the "glimpses".

4. Experiment 2

4.1. Subjects and experimental setup

The subjects in this experiment were 111 students; 101 attended courses in engineering, of which 16 declared musical experience of at least two years, like playing a musical instrument or an involvement in some audio engineering task; 10 attended a musical degree course. None of the students had previous experience with psychoacoustics tests. For technical reasons the experiment was held in two different rooms, I and II, in nine groups of 6–14 listeners. Engineering students used room I, while musical degree students used room II. Both rooms were of similar size (45 and 55 m^2 respectively) and similar acoustical characteristics. In room II the reverberation time (RT20) was: 0.9 s at 128 Hz, 0.75 s at 1 kHz and 0.45 s at 4 kHz. The stimuli were played through one loudspeaker. The test setup in Room I: PC computer with Prodif 88 digital audio interface, Swissonic DA96 D/A converter,

Soundcraft Spirit 16/8 mixer and Genelec 1038A studio monitor; in Room II: PC computer, Digi 002 audio interface and Mackie 626 monitor. The listening level was set at about 80 dB according to the preference of listeners in each of the groups. All sound samples were normalised to the same RMS value. To minimize the effect of the place of listening, the subjects in large (8 and more) groups were asked to change their places randomly after having heard half of the samples.

4.2. Experimental procedure

Pilot experiments with a loudspeaker indicated that the proportion of correct recognitions did not exceed chance performance for all 12 sets in individual cells versions, in contrary to clustered cells versions, where significant differences were met. Therefore all clustered cells versions of Table 1 plus the individual cells version of the "all_5" set were chosen for the main experiment. This choice was supported by the results of experiment 1, where individual cells mode produced substantially lower detectability than clustered cells mode. Full 7 s long excerpts were used. The paradigm of a trial followed the ITU BS.1116-1 recommendation (ITU, 1997) for the evaluation of subtle differences in the quality of audio signals or equipment. An AXY test was used. The trial consisted of three observation intervals (A, X and Y), where A was a reference and was repeated in either X or Y. The other component of the X, Y pair was a stimulus different than a reference, i.e. it was a triple stimulus, hidden reference test. The subject's task was to indicate whether the interval X or Y contained the different signal. The original mix was used as a reference and assigned to the interval A. The listeners were given score sheets with the instructions. To make the task easier for subjects the sequences in intervals in a test trial were named AAB and ABA. The experiment started with a training run of all sets, then the main test with 13 test sequences followed. The sequence of presentation was as follows: double alert signal – AAB sequence with 500 ms breaks between intervals – single alert signal – AAB sequence (repeated) – 5 s for giving the answer. The reason for repeating was to help the listener in case he/she was not sure after just one hearing. In half of the trials the sequence was ABA. As listeners auditioned the test sequences in groups, all members of a group heard the same sequence. The sequence of sound sets (13 sets) between groups was randomised, and so were the sequences for a particular set (for a given set ABA sequence was used for half of the groups and AAB for the other half). For each of the 13 sets the score sheet contained a field to be filled with the recognised sequence. The correct sequence (either AAB or ABA) indicated that a listener recognised correctly the stimuli, i.e. recognised the difference between A and B.

4.3. Results and discussion

The results have been evaluated by two methods. One is used in some listening tests of audio equipment. The results were treated as categorical data (2 categories) and their significance level was determined from a binomial distribution. The other is used in psychophysics. A percentage of listeners who correctly recognised was treated as the percentage of answers for one listener. The decision audible/inaudible was then based upon the threshold of 75%.

In only four out of 13 sets was the difference recognized correctly by a significant ($p < 0.05$) proportion of listeners. In other sets, the results were far below significance. The results for the four significantly recognised stimuli are presented in Table 3. The p – value has been determined from a one-tailed binomial distribution.

Table 3. Statistically significant correct recognitions in experiment 2, all listeners. Alternative evaluation based on the psychophysical threshold of 75% is given in the right column.

Set of instruments	Percent correct recognitions	p – value	Effect audible according to 75% rule
dr-g-sn-sx	61	< 0.01	no
dr-sn-sx	63	< 0.01	no
dr-g-sn	73	< 0.001	no
g-sn	73	< 0.001	no

No common property was found in these four sets. Informally, listeners commented that they recognised differences on the basis of small artifacts, different in each of the sets.

The results for 26 musically experienced listeners (16 from engineering courses declaring at least two years of musical experience, and 10 from music courses) are given in Table 4. The paired-data t test was performed (using four sets occurring in both Tables 3

Table 4. Statistically significant correct recognitions in experiment 2, musically experienced listeners only. Alternative evaluation based on the psychophysical threshold of 75% is given in the right column.

Set of instruments	Percent correct recognitions	p – value	Effect audible according to 75% rule
dr-sn-sx	69	< 0.05	no
dr-g-sn-sx	77	< 0.01	yes
dr-g	77	< 0.01	yes
dr-g-sn	77	< 0.01	yes
g-sn	85	< 0.001	yes

and 4) to find whether musically experienced listeners were more sensitive to differences between versions of sound mixes. The result was positive at $p < 0.025$ (t-value = 3.45, df = 3). Although the difference in listening rooms and equipment were unlikely to affect the results, the results of students of music courses were not treated separately, to avoid the effect of this theoretically confounding factor. Creating a group for "musically experienced" listeners instead, of which 10 listened in room I and 16 in room II was meant to average out possible room effects.

The relation between percentages of the energy and area retained and the rate of recognition of RSO processing in this experiment is less pronounced than in experiment 1, but the same rule can be observed. The average energy retained in the clustered cells mode computed from to Table 1 is 72.2%, while the average energy in significantly recognised sets in Table 3 is 68.5%, and in Table 4 (musically experienced listeners): 66%. In the case of the area retained, the analogous percentages are: 54.6%, 53.5% and 49.3%.

Neither of the two stimuli assessed in experiment 1 as detectable (mean $d' > 1$) were detected in experiment 2, even by musically experienced listeners, indicating that the perceptual effect of RSO is substantially weaker when auditioned over the loudspeaker.

5. Conclusions

The following conclusions can be drawn from this work:

1. The perceptual effect of the operation of artificial removal of spectro-temporal overlap was imperceptible in one of four experimental conditions in experiment 1. This condition was easy to meet: a mix of two sound sources and individual cells processing, which can be considered a natural option for RSO. Therefore, the general conclusion is that the effect of RSO can be imperceptible.
2. Clustering of cells made the effect perceptible in both of the stimuli investigated in experiment 1.
3. In loudspeaker listening (experiment 2), the range of conditions in which the effect of RSO was imperceptible was considerably wider than in headphone listening. The effect was not perceived in 9 out of 13 stimuli investigated (69%), although most of them were of "clustered cells" type, found as perceptible in headphone listening.
4. The detectability of RSO processing increases with removing more energy and effective t-f area.
5. The effect of RSO in its imperceptible condition as found in this work indicates that the multiple looks/glimpses hypotheses hold in the perception of musical instruments. The results also indicate that "glimpses" are quite narrow in frequency, in the order of a fraction of the CB.

Appendix. The computation of relative width of t-f regions

In order to concentrate the analysis on the frequency widths of t-f components, the regions were assumed as one-cell wide vertical strips of cells in the t-f plane, consisting of contiguous cells belonging to one instrument (any strange cell broke the strip). Individual time frames were analysed, therefore it did not matter whether a strip was isolated in time, or was attached to any cells of the same instrument on either side of the strip. This approach was different from the assumption in (Cooke, 2006) that a glimpse (i.e. a region) contained all cells connected by being a part of the four-neighbourhood of any other element in the t-f region.

The aim was to count t-f regions of a similar logarithmic bandwidth. The computation was carried out in the range from 1270 to 9500 Hz. The lower frequency was chosen so that the relative bandwidth analysed was not wider than about 15% of a CB. Higher frequencies were not included because of considerable share of background noise in that band. The number of contiguous cells was counted separately in each of 12 CBs in the analysed range. The appropriate margin was included in the algorithm, so that wide regions exceeding limits of CBs were not broken and their whole width was counted. The results are approximate because no perfect alignment between fractions of CBs, limits of CBs and multiples of the cell's width could be obtained. The counting was carried out for both instruments in the pair and the results were averaged.

Acknowledgment

This study was partly supported by grant number R02 0030/2009 from the National Centre for Research and Development (Polish).

References

1. Barker J., Cooke M. 2007, *Modelling speaker intelligibility in noise*, Speech Communication, **49**, 402–417.
2. Bregman A.S. (1990), *Auditory Scene Analysis*, MIT Press, Cambridge.
3. Brungart D.S., Chang P.S., Simpson B.D., Wang D.L. (2009), *Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers*, J. Acoust. Soc. Am., **125**, 4006–4022.
4. Cooke M.P. (2006), *A glimpsing model of speech perception in noise*, J.Acoust. Soc. Am., **119**, 1562–1573.
5. Fastl H., Zwicker E. (2007), *Psychoacoustics – facts and models*, Springer – Verlag, Berlin Heidelberg.
6. Gescheider G.A. (1997), *Psychophysics: The Fundamentals*, Lawrence Erlbaum Associates.

7. Howard–Jones P.A., Rosen S. (1993a), *The perception of speech in fluctuating noise*, Acustica, **78**, 258–272.
8. Howard–Jones, P.A., Rosen S. (1993b), *Uncomodulated glimpsing in 'checkerboard' noise*, J. Acoust. Soc. Am., **93**, 2915–2922.
9. ITU (International Telecommunication Union) (1997), *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*, Recommendation BS.1116-1.
10. Kelly M.C., Tew A.I. (2002), *The continuity illusion in virtual auditory space*, Proc. 112th AES Conv. Preprint 5548.
11. Kelly M.C., Tew A.I. (2003), *The significance of spectral overlap in multiple-source localization*, Proc. 114th AES Conv. Preprint 5725.
12. Kingdom F.A.A., Prins N. (2010), *Psychophysics: A Practical Introduction*, Academic Press, London.
13. Kleczkowski P. (2002), *Acoustic Signal Expansion in Multiple Trigonometric Bases*, Acta Acustica united with Acustica, **88**, 526–535.
14. Kleczkowski P. (2005), *Selective Mixing of Sounds*, 119th AES Conv., New York, Preprint 6552, October 2005.
15. Kleczkowski P. (2008), *Selective mixing of a symphony orchestra recording*, Archives of Acoustics, **31**, 4 (Supplement), 91–99.
16. Kleczkowski P., Plewa M., Pluta M. (2010), *Masking a frequency band in a musical fragment played by a single instrument*, Acta Physica Polonica A, **119**, 991–995.
17. Kleczkowski P., Pluta M. (2012), *Understanding concurrent speech is not impaired by removal of spectro-temporal overlap*, Acoustics 2012 Conference, Hong Kong.
18. Lu Y., Cooke M. (2008), *Speech production modifications produced by competing talkers, babble, and stationary noise*, J. Acoust. Soc. Am., **124**, 3261–3275.
19. Malvar H.S. (1992), *Signal Processing with Lapped Transforms*, Artech House, London.
20. Moore B.C.J. (2003), *An Introduction to the Psychology of Hearing*, Academic Press, London.
21. Moore B.C.J., Glasberg B.R., Plack C.J., Biswas A.K. (1988), *The shape of the ear's temporal window*, J. Acoust. Soc. Am., **83**, 1102–1117.
22. Viemeister N.F., Wakefield G.H. (1991), *Temporal integration and multiple looks*, J. Acoust. Soc., **90**, 858–865.
23. Wang D.L., Brown G.J. (2006), *Fundamentals of computational auditory scene analysis* [in:] Wang D.L., Brown G.J. [Eds.], *Computational auditory scene analysis: Principles, Algorithms, and Applications*, IEEE Press/Wiley-Interscience., Hoboken NJ, pp. 1–44.