ARCHIVES OF ACOUSTICS
31, 4 (Supplement), 409–416 (2006)

SOME EXPERIMENTS ON EXTREME MASKING

P. KLECZKOWSKI

AGH University of Science and Technology
Department of Mechanics and Vibroacoustics
Al. Mickiewicza 30, 30-059 Kraków, Poland
e-mail: kleczkow@agh.edu.pl

(*received June 15, 2006; accepted September 30, 2006*)

Some experiments are described, in which the structure of the information received by the human ear is made to be comparable to the structure of the information received by the eye. The appropriate processing of the acoustic signals is performed in the time-frequency plane. This operation can also be seen as an artificial, extreme masking, where there is only one masker in each region of the time-frequency plane and all other sounds are entirely masked. The experiments were performed on audio signals. It was found that the ear perceives little difference between the original sounds and the processed sounds.

Key words: psychoacoustics, auditory scene analysis, audio engineering.

1. Introduction

The hearing system receives the sum of acoustic pressures generated by various sources. The entire sum is not perceived however, because of masking. Masking with multiple simultaneous sources is a complex phenomenon, less known and understood than masking of a maskee by a masker. In this work we investigate the case when sounds of comparable levels from some sources arrive at the ear. This is usually the case when the music is played. The appropriate way to analyse masking then is by investigating the appropriate time-frequency distributions. In some regions of the time-frequency plane the elements of one of the sounds will mask the elements of the other sounds. In other regions none of the sounds is dominating enough to mask the others, and consequently two or more sounds are heard simultaneously.

Some similarities to the above structure of information can be found in the information received by the eye. However, the occlusion (corresponding to masking) in signals arriving to the eye is always complete. In Fig. 1 we see three objects: a dolphin, a ship with its shadow, and a mountain lake. No part of the dolphin, which is the closest object, is occluded. Large parts of the ship are occluded by the dolphin, and large parts of the mountains and the lake are occluded by either the dolphin, the ship, or both of them. At no part of the picture more than one object can be seen. We can only see one of them

at a given place in the picture. This rule for the vision would be broken only if one or more of the objects were semi-transparent.

Fig. 1. A simulation of a view of three objects: a dolphin, a ship with its reflection in water and a mountain lake.

This work describes some experiments, where the scheme of occlusion typical of the pictures received by the eye is implemented in acoustic signals reaching the ear. This procedure can be compared to a specific, artificial sort of masking and will be referred to by the author as “extreme masking”.

The purpose of this work is to investigate the general effect of the presented processing on the human perception of sound.

The only works known to this author which can be considered close to this issue had been published by KELLY and TEW [2–4]. However, the scheme they used was different, and can not be compared to extreme masking. Also the objective of their work was different, as they investigated the possible effects of their operation on the stereo image.

2. The idea of extreme masking

The idea of extreme masking is illustrated in Fig. 2. The time-frequency spectrum of combined sounds of a saxophone and a synthesizer is shown. Both sounds have similar sound pressure level. Black colour indicates elements belonging to the spectrum of a saxophone sound, while grey indicates elements belonging to the spectrum of a synthesizer sound. The two spectra have been combined by the comparison of energy content in each of the time-frequency cells. For each cell in the time-frequency plane, only the contents belonging to the stronger of the two instruments in that cell is present. There is a direct correspondence between this plot of two sounds and the picture in Fig. 1 and its scheme of occlusion. In both cases, in each point of both of the planes (the time-frequency plane and the plane of the picture) there is either an element of one of the two sounds, or an element of one of the three visual objects. In no point of either of the planes there are superimposed sounds nor superimposed visual objects.

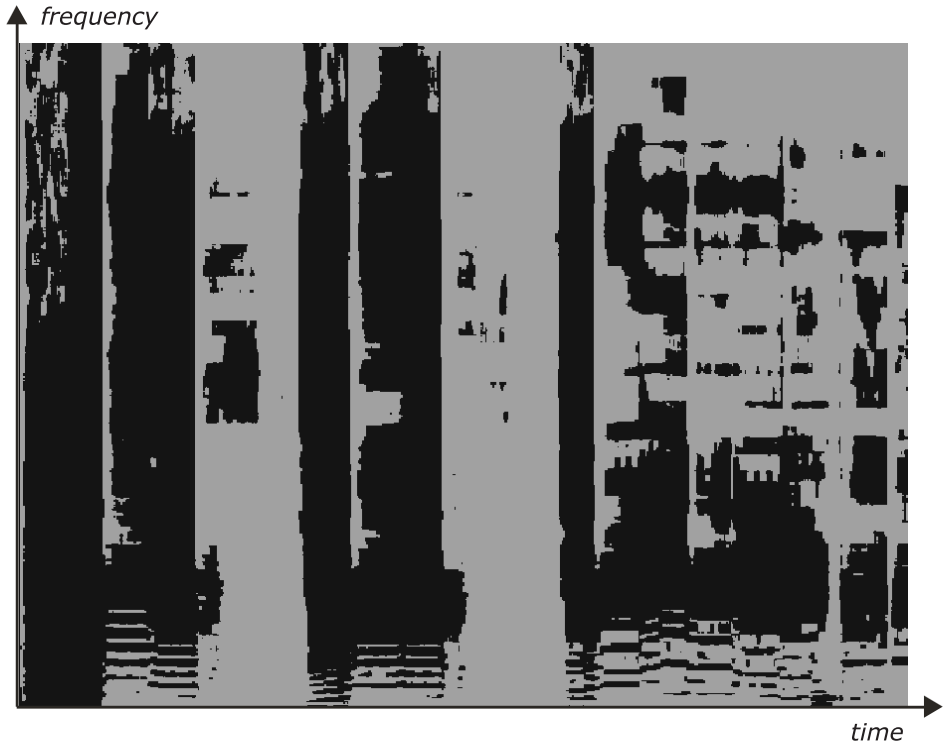

Fig. 2. An example of the result of simulation of extreme masking, performed for two sound sources: a saxophone (black area) and a synthesizer (grey area).

As an effect of the processing described above and shown in Fig. 2, in each point of the time-frequency plane one of the instruments completely masks the other, or occludes it by some analogy with vision and Fig. 1. This is why the term “extreme masking” is proposed.

Extreme masking is a simulation of a specific acoustical effect, but this effect should not be considered a sort of masking.

3. The masking versus extreme masking

There are substantial differences between the masking and the effect of extreme masking.

The experiments reported here were conducted on audio signals. It is not straighforward to qualify masking of an audio signal by another audio signal to one of the widely documented models of masking, but the following differences can be pointed out.

1. In simultaneous masking of a tone by a narrow band of noise the masker must exceed the level of the test tone by 15–20 dB [7]. In simultaneous masking of the test tone by the harmonic complex the masker must exceed the test tone by more than 10 dB [9]. In our example, when the saxophone and the synthesizer play together, in large parts of the time-frequency plane they are heard simulta-

neously, as there is no masking. In extreme masking, any value of the difference in levels greater than 0 cause masking, so masking occurs everywhere in the time-frequency plane.
2. No temporal masking is included in the effect of extreme masking.
3. In simultaneous masking of a tone by a tone side effects occur like beating and difference tones, resulting from nonlinearities in the ear. Elements of masking of a tone by a tone are present in audio signals and no such side effects occur in extreme masking.

Another phenomenon related to extreme masking is continuity illusion. On the basis of this phenomenon, presented widely in BREGMAN's book [1], one may expect, that switching off of short elements of signals, in the order of no more than a couple hundred of miliseconds, may be not perceptible. According to [1] this requires that a couple of additional conditions are met, of which the most important is the requirement that there is a meaningful difference between the levels of signals (the dominating signal and the signal which is switched off).

The effect presented in this work considerably extends beyond that of the continuity illusion. No limits at all have been assumed for extreme masking, it is attempted to reconstruct the structure of a picture shown in Fig. 1, and consequently some sounds may become entirely eliminated.

4. The simulation of extreme masking

The simulation of extreme masking of elements of sounds in the time-frequency plane, presented in the simplest case of two sounds in Fig. 2, can be implemented for any number of sound sources.

In order to implement the effect it is necessary to have separate recordings of sound objects. The author has used a multitrack recording in all of his experiments. The following operations need to be performed:

1. The conversion of each of the signals (tracks of a multitrack recording) from the time domain into the time-frequency domain.
2. The comparison of the time-frequency spectra of all signals with choosing, for each of the elements of the time-frequency plane, the signal with the highest energy within that element. This operation is illustrated in Fig. 3.
3. For each of the signals, cancelling all of its elements in the time-frequency plane, except those, which in operation 2 have been chosen as having the highest energy. The cancelling of elements is the actual simulation of extreme masking.
4. The summing of such processed signals in the time-frequency domain.
5. The conversion of such obtained sum of signals into the time domain. Alternatively, it is possible to convert separate signals after the operation of cancelling (operation 3) into the time domain and summing them in this domain.

The elements of the time-frequency plane, shown in Fig. 3, can be elementary cells of a time-frequency representation, having the minimum area for a given representation. This area is limited by the uncertainty principle. The elements can also be formed from a group of such cells.

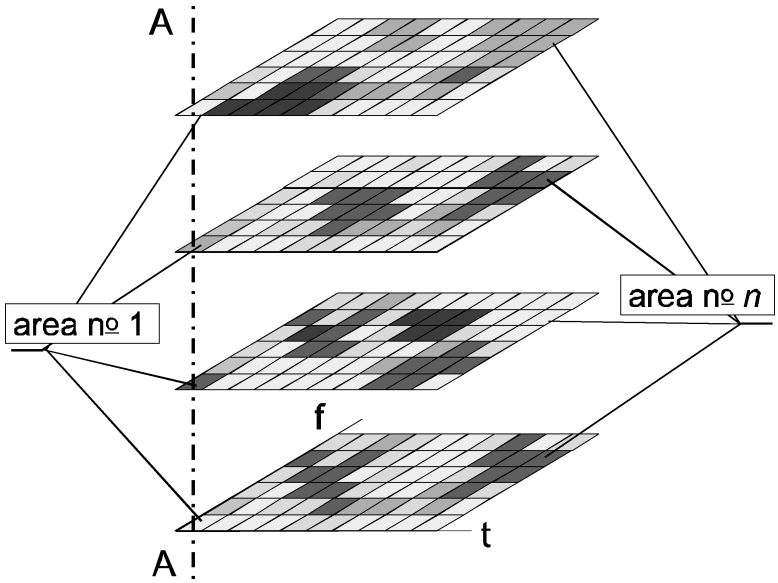

Fig. 3. In the four planes, the examples of time-frequency representations of four sounds are shown. The planes have been divided according to the common rectangular grid into elements. The energy of a signal within each element is indicated by grey scale. The comparison of energies (operation 2) is performed for a particular element of the plane, e.g. along the A-A line.

The output signal containing the results of the simulation of extreme masking can only be reproduced by a single audio channel. In order to simulate it in sounds arriving from different directions, the operations 1–5 should be performed independently for all signals assigned to an audio channel representing a given direction. This has been implemented by the author for the two channel (stereo) case.

5. Shaping the areas in the time-frequency plane

If the elements of the time-frequency plane upon which the operation no. 2 is performed are the elementary cells of a given representation, then the result of operations 2–4 is that the entire plane is covered by a large number of small, scattered areas. This is partly shown in Fig. 2, but in fact the result of closely following of rules 2–4 would be even more jugged, as a smoothing algorithm was used in computing the pattern of Fig. 2.

When the spectra of sounds in the time-frequency plane are jugged, then rapid switching between the signals occur and this results in nonlinear distortion of the result of the simulation. Although this distortion is hardly perceived in the output signal, it should be avoided, in order to investigate the effect of extreme masking alone.

Clustering of the cells into groups of rectangular shape reduces the number of switching between the signals and hence reduces distortion, but it has a serious drawback of loosing precision of determining the borders between the neighbouring clusters.

The clusters of cells may have shapes different than rectangular, rather irregular, and the plane may be covered with clusters of different shapes. If the shapes of the clusters are different than rectangular, then between the operations no. 1 and no. 2 of Sec. 3 one more operation is needed:

- the division of the plane into areas, depending of the spectra of particular signals.

This operation is the most difficult of all and requires advanced algorithms. The appropriate division of the plane should precisely match the borders of the sound objects, in order to preserve features of sounds which are important for their perception. However, the areas should be possibly coherent and their borders should be smooth. These both requirements are contradictive so this work requires numerous experiments in order to find values of parameters for the best perceptual quality of the output signal.

Simple and efficient algorithms have been developed on the basis of the rules of neighbourhood. They consist in assigning to each of the elementary cells a substitute value of energy. This value is equal to the sum of energies in a number of cells in its neighbourhood plus its own energy. Next, according to the rule of operation no. 2 (Sec. 3) the energies assigned to each of the cells at the particular location in the plane are compared. The algorithms using neighbourhood rules tend to make the clusters coherent and they smooth out the borders between them. In Fig. 4, basic rules of neighbourhood are shown: Moore's rule and von Neumann's rule [6].

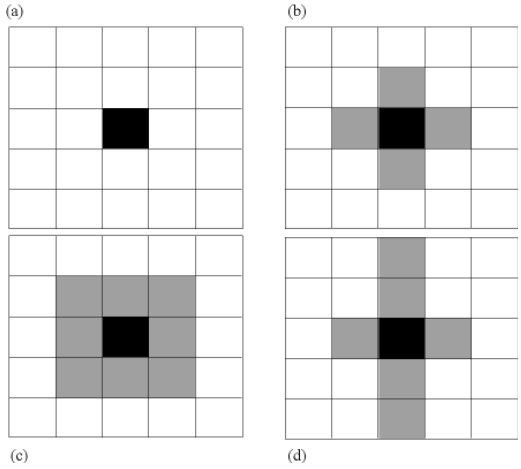

Fig. 4. (a) – a single cell, (b) – neighbourhood according to von Neumanna rule, (c) – neighbourhood according to Moore's algorithm of first order, (d) – neighbourhood according to an asymmetric von Neumann algorithm.

The methods of clustering are under further development. In [6] a method based on the two-dimensional Fourier transform and two methods based on probabilistic Monte Carlo approach [8] are presented. The perceptual differences between the methods of clustering are subtle so the experimental work needed for their evaluation is extensive and no credible results may be presented yet. Initial experiments tend to indicate that the Monte Carlo method of minimising the perimiter of clusters yields good results.

Its idea is shown in Fig. 5. It consistst in minimizing the perimeter of all clusters in the plane. The starting point is the result of the operation no. 2 of Sec. 3. Then, the algorithm proceeds as follows:

1. Select a pixel at random.
2. Change the attachment of the pixel into an attachment to a neigbouring cluster.
3. Compute the change in the length of the boundary.
4. The change will be accepted with the probability:

$$1 - \exp\left(-\lambda(B_{\text{after}} - B_{\text{before}})\right), \tag{1}$$

where B_{after} is the boundary length after the change and B_{before} is the boundary length before the change. In fact, we do not need to compute the total boundary length each time, but only the change due to the new value of the pixel. The parameter λ controls the rate at which the pixels are removed.

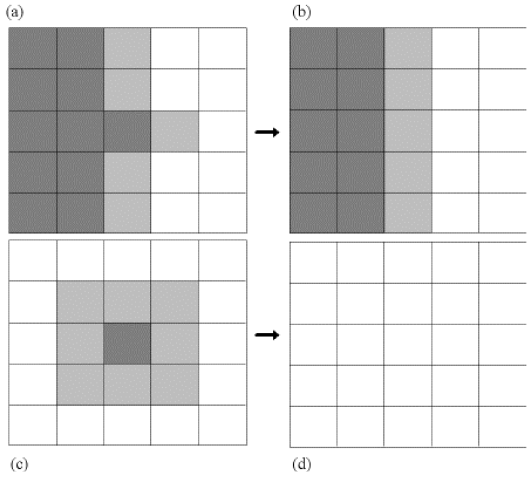

Fig. 5. The idea of a Monte Carlo method minimising the perimiter of a cluster. (a) – a part of a perimeter indicated by light grey, (b) – the same part after minimisation, (c) – the perimeter of a single cell, (d) – a meaningful shortening of the perimiter after an isolated cell has been cancelled.

6. Conclusions from listening tests

Listening tests have been carried out with the audio materials prepared according to the rules of Sec. 3, and also extended by techniques mentioned in Sec. 5 and described in [5] and [6]. The following general conclusions can be formulated:

1. Depending on the material and algorithm used, the effect of extreme masking is just noticeable or hardly noticeable.
2. The intensity of perception decreases with the number of sound sources.

Further conclusions:

When the areas in the time-frequency plane are shaped according to the appropriate techniques and further operations presented in [6] are used, the extreme masking has the following features, which are advantageous in audio engineering:

1. The sound is more detailed.
2. If the input signals contain microphone cross-talk, then the extreme masking attenuates them considerably.

The sound obtained with the algorithms currently used by the author has the following other features:

3. The sound is brighter.
4. The sound is less “warm” than the sound mixed without the simulation of obscuring.

7. An attempt to explain the effect of extreme masking

The proven fact, that extreme masking does not deteriorate the quality of the perceived sound needs explanation, as the effect extends beyond the phenomena of masking and continuity illusion. Further research is needed, but a very tentative hypothesis might be the following. Under specific circumstances the ear may include narrowband elements of sounds both to a sound to which they belong, and to other sounds, where they do not belong, but which have been deprived of their own elements in the same narrow bands of frequencies and in the same segments of time. The justification of the above hypothesis is that an area of the time-frequency plane corresponds to a narrowband segment of sound in the time domain. From the point of view of the ear analysing one particular sound, some narrowband elements of the original sound have been substituted by alien narrowband elements. However, these alien segments differ mainly in their phase, as there are means to minimise the differences in amplitude [5]. Thus, the process of subjective assimilation of the alien narrowband elements to sounds deprived of their own elements in corresponding areas of time-frequency plane may take place.

References

[1] BREGMAN A. S., *Auditory scene analysis*, MIT Press, Cambridge 1990.

[2] KELLY M. C., TEW A. I., *The continuity illusion in virtual auditory space*, 112-th Convention of the Audio Engineering Society, Munich, May 2002, preprint 5548.

[3] KELLY M. C., TEW A. I., *The continuity illusion revisited: coding of multiple concurrent sound sources*, Proc. 1-st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002), Leuven, Belgium, November 2002, 9–12.

[4] KELLY M. C., TEW A. I., *The significance of spectral overlap in multiple-source localization*, 114-th Convention of the Audio Engineering Society, Amsterdam, March 2003, preprint 5725.

[5] KLECZKOWSKI P., *Selective mixing of sounds*, 119-th Convention of the Audio Engineering Society, New York, October 2005, preprint 6552.

[6] KLECZKOWSKI P., KLECZKOWSKI A., *Advanced methods for shaping time-frequency areas for the selective mixing of sounds*, 120-th Convention of the Audio Engineering Society, Paris, May 2006.

[7] MOORE B. C. J., *Masking in the human auditory system*, [in:] *Collected Papers on Digital Audio Bit-Rate Reduction*, Audio Engineering Society, 1996.

[8] OTTEN R. H. J. M., VAN GINNEKEN L. P. P. P., *The annealing algorithm*, Kluwer, Boston 1989.

[9] ZWICKER E., FASTL H., *Psychoacoustics, facts and models*, Springer-Verlag, Berlin 1990.