Audio Engineering Society

Convention Paper 6718

Presented at the 120th Convention
2006 May 20–23 Paris, France

Advanced Methods for Shaping Time-Frequency Areas for the Selective Mixing of Sounds

Piotr Kleczkowski[1], and Adam Kleczkowski[2]

[1] AGH University of Science and Technology, Dept. of Mechanics and Vibroacoustics, 30-059 Krakow, Poland
kleczkow@agh.edu.pl

[2] Selwyn College, University of Cambridge, Cambridge, CB3 9DQ, England
adamk@mathbio.com

ABSTRACT

The Selective Mixing of Sounds is a large procedure, a large part of which is conceptually challenging and has not been developed previously. This is a method of determining areas in the time-frequency plane. It has a major effect on the overall quality of the sound. In this paper we propose and compare a range of appropriate algorithms. We begin with a simple two-dimensional running mean combined with a rule selecting the track characterised by the maximum energy, followed by a low-pass filter based on the 2-dimensional Fourier transform. We also propose two novel methods based on the Monte-Carlo approach, in which local probabilistic rules are iterated many times to produce the required level of smoothing.

1.INTRODUCTION

The Selective Mixing of Sounds is a new approach to mixing, proposed in [1]. The original paper mainly focused on a new effect found in psychoacoustics: the removal of large parts of musical tracks in the time-frequency domain may be not perceived in the mix at all. When carefully exploited, this rule offers several advantages for the mixed sound. The main advantage is that the sound can be made more detailed and this has been confirmed by the vast majority of participants of numerous listening tests.

The effect described in [1] has some relation with an experiment first conducted by Dannenbring in 1976, and further work, analysed by Bregman in [2]. This family of effects is referred to as “the illusion of continuity”. Kelly and Tew [3], [4], [5] investigated the effect of removing fine spectral detail in the regions of spectro-temporal overlap between two sound sources on their localisation. They found that it was possible to remove the spectro-temporal components of a sound as long as their level was at least 10 dB to 15 dB below the components belonging to the other sound.

In the Selective Mixing of Sounds, we compare the time-frequency spectra of any number of individual tracks of audio to be mixed to one output channel and remove parts of these spectra according to local energy levels. The goal is to improve the sound of the mix.

The method of comparison of local energy levels is also different than the one used in [3], [4] and [5]. Kelly and Tew compared the cells of a fixed grid in the time-frequency plane. In [1] the idea was to find the appropriate regions in an intelligent, context – dependent way, so that the perceptually important elements of each of the tracks are preserved entirely. The details are presented in this paper.

The processing starts with performing time-frequency analyses of all tracks to be mixed. Then, the patterns of all tracks are analyzed and combined in the time-frequency plane in order to select areas of high and low importance. Shaping of areas and the decisions of assignment to either of the categories must take the contents of all of the tracks into consideration. The next stage of processing consists in removing of all areas of low importance from all of the tracks. The last stage is the mixing of the remaining areas of high importance.

The rule of the shaping of areas is important for the overall quality of sound obtained. The goal of the effective shaping-assignment procedure is to preserve all characteristic shapes of time-frequency patterns of a given instrument, as long as they can be perceived in the mix. There is a contradiction between the shapes of the areas preserving the characteristic details of a given instrument and these shapes being smooth. Smoothness increases the overall clarity of the mix, but when the shapes are too smooth some details may be lost resulting in perceptible distortion. It is not possible to determine a priori which of the mathematical tools for computing the shapes of the areas will provide the best balance between smoothness and detail, therefore we investigate three ways of approaching it.

In the simplest approach, the separation of areas is done separately for each of the tracks and the resulting combination of patterns is smoothed in the time-frequency plane. The areas are treated as a two-dimensional pattern to which different global smoothing techniques are applied. We begin with a simple two-dimensional analogue of a running mean, analogous to a low-pass filtering in the signal domain. As an alternative, we consider another low-pass filter, based on the 2-dimensional Fourier transform of the pattern. The filter can be asymmetric in two dimensions, reflecting differences in properties of the time and frequency axis spanning the time-frequency plane.

The advantage of this method of smoothing lies primarily in its simplicity of a concept and implementation and a low computational cost. However, the lack of flexibility is a major drawback, particularly for preservation of individual track properties. We therefore propose a range of novel methods, based on an iterative Monte-Carlo approach. While computationally more expensive, these methods are more flexible and can combine the shaping assignment and smoothing into a single step. The approach focuses here on a single time-frequency event and uses a set of deterministic or probabilistic rules to assign to it a particular track. The selection depends not only on the properties of all tracks at the event, but also on its neighborhood. The rules are then applied sequentially to all events, either in a systematic or random order. The procedure is then repeated until the required level of smoothing is achieved. In this way, the selection of a dominant track for any given time-frequency event is based on a large number of events. This allows preservation of some characteristic details of an instrument, while selecting the most important signal for each area of the time-frequency plane.

2. FILTERS

Mixing and filtering of signals is a non-linear process in which we attempt to reduce the information from several tracks to a single track. In our approach, we first construct maps of energy as a function of time (horizontal axis) and frequency (vertical axis), for each track. The next step is to combine the individual maps and we consider two outcomes of this process. The first is an *occupancy map* showing which portions of the time-frequency domain are assigned to a given instrument. Secondly, the *energy distribution* is plotted over the distribution map. In this paper we are primarily concerned with a construction of the *occupancy map*. We assume that for the distribution map, each pixel can take integer values coding the dominant signal at each combination of time and frequency value.

In a simplest case, a signal with the highest energy in the given pixel is chosen (LMR: *local maximum rule*). This is illustrated in Fig.1. There are several disadvantages of this approach: Two tracks might be very close to each other in the energy spectrum at a given time. Because the energy spectrum can be locally irregular, the combined effect leads to a random selection and switching on and off of tracks. In addition, there is no preservation of a characteristic shape of how different signals develop in time.

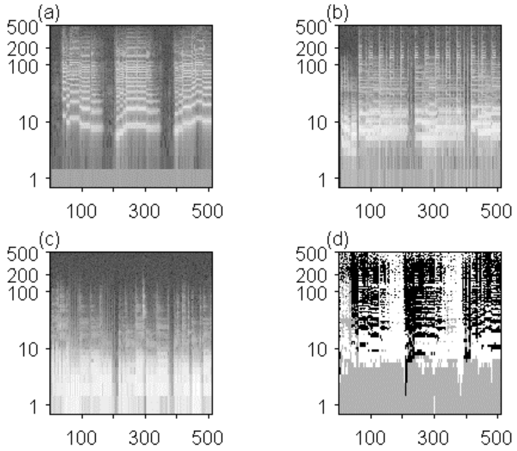

Figure 1. (a-c) shows an example of 512ms of a signal from 3 instruments: (a) saxophone, (b) guitar and (c) bass. Horizontal axis – time [ms], vertical axis – frequency [number of the frequency bin]. Frequency is shown using a logarithmic axis; light areas correspond to high values of energy, whereas dark areas to low values of energy. A map constructed using a local maximum rule, is also shown. Regions dominated by bass are shown in grey, by guitar in white and by saxophone in black.

2.1. Deterministic neighborhood rules

What is missing in the picture above is non-local information. In particular, we want to include information about the energy in the neighborhood, both in time and in the frequency domain. Figure 2 shows how information from neighborhoods of various size and shape can be combined with the *maximum rule*: To determine occupancy of the central pixel (marked in black in figure 2), we consider all its neighbors (marked in grey) in either the symmetric von Neumann neighborhood (b), a Moore neighborhood (c) or in an asymmetric von Neumann neighborhood (d).

For a given neighborhood structure, the value of the central pixel is determined by calculating the average energy for each signal in the neighboring cells and then choosing the signal with the highest energy. Thus, if a saxophone has the highest average energy over the neighborhood, the central pixel will be marked in black. In figure 3 we show results for the von Neumann neighborhood consisting of 4 nearest neighbors, a second-order Moore neighborhood (as in (c) but including 24 neighbors) and in figure 4 the asymmetric von Neumann neighborhood, as in (d) but including 10 pixels on each side in the vertical (frequency) or horizontal (time) direction.

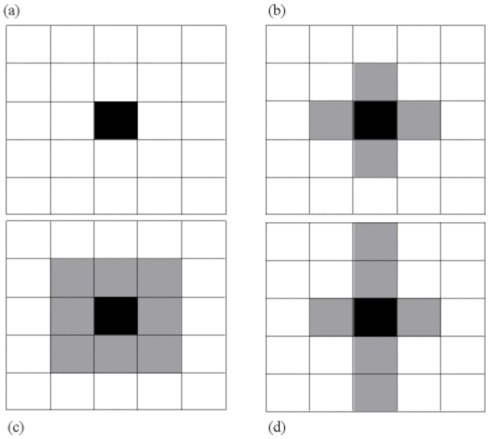

Figure 2. Concept of the neighborhoods. (a) shows the central pixel (marked in black in (a-d)), (b) represents a von Neumann neighborhood, (c) a first-order Moore neighborhood and (d) an asymmetric von Neumann neighborhood.

2.2. 2-dimensional Fourier transform

In the above method we can only realistically include small-sized neighborhoods. A more flexible approach is provided by treating the *occupancy map* obtained by applying a LMR as a two-dimensional image and subsequently performing filtering on this image. In this approach, an original energy distribution is first replaced by arbitrary numbers representing each instrument (as described above). In our example, each pixel where saxophone has a largest energy of its signal is represented by number 1; each pixel where guitar is dominating is allocated number 2; bass guitar is then allocated number 3. The resulting occupancy map is shown in figure 1d, and is characterized by a large scatter of points. When a 2-dimensional Fourier spectrum [6] is computed for such a pattern, it is dominated by a central peak reflecting the final size of the signal, with additional contributions from some low-frequency (2-dimensional) signal corresponding to large-scale patterns where one instrument is dominating.

The 'tail' of the spectrum corresponds to small-scale oscillations. We used a Gaussian (2-dimensional) filter to separate out the domains. The width of the filter can be different in the horizontal dimension (corresponding to a 'real' time) than in the vertical dimension (corresponding to a 'real' frequency). Figures 5a and 6a show examples of the filter used in our calculations. In figure 5, more information in the vertical direction (frequency in the original occupancy map) is retained in the filtered signal than in the horizontal direction (time in the original occupancy map). In other words, more smoothing is achieved in the horizontal direction (time) and more detail in the vertical direction (frequency). The signal in figure 6 retains more information in the horizontal direction, whereas the vertical direction is more smoothed. After the filter is applied, we convert the signal back to the original variables by applying an inverse Fourier transform.

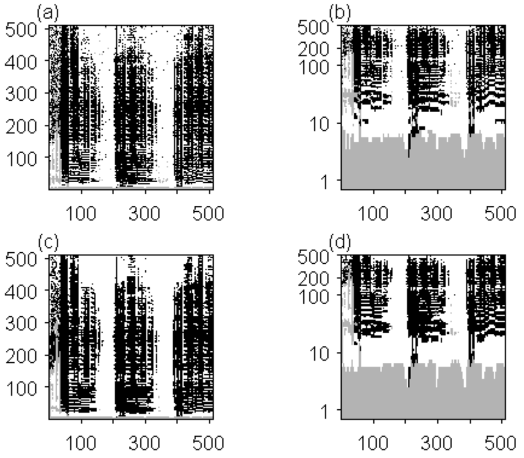

Figure 3. Maps of signal distribution: Symmetrical neighborhoods: von Neumann first order neighborhood (a) and (b), and the second order Moore neighborhood (c) and (d). Regions dominated by bass are shown in grey, by guitar in white and by saxophone in black. Logarithmic scale for the vertical (frequency) axis is used in (b) and (d) to show low frequencies dominated by bass. The scale of the vertical axis denotes the number of the frequency bin.

If no information was rejected by the filtering process, the distribution of values in the filtered signal will be at discrete points, corresponding to 1 (saxophone), 2 (guitar) and 3 (bass guitar). However, filtering distorts this property of the distribution and the resulting pattern takes values from a continuum of numbers between 0 and 3. We convert this distribution back into the original set of labels {1,2,3} by applying discrete binning. The conversion of a smooth distribution into a discrete set of labels can be used to manipulate the signal further. Similarly, an order of the instruments can be used to manipulate relative contributions of instruments.

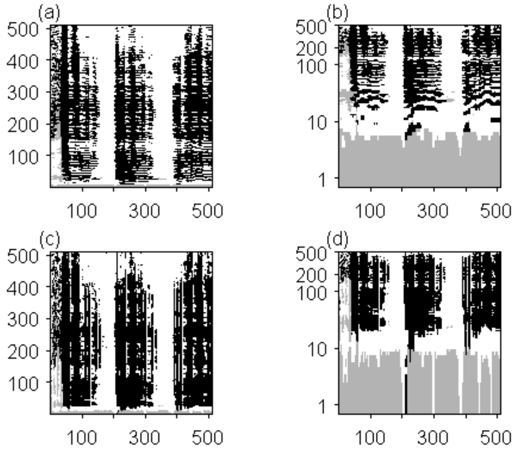

Figure 4. Asymmetric von Neumann neighborhoods: Extended in the horizontal direction (a) and (b), and in the vertical direction (c) and (d). Regions dominated by bass are shown in grey, by guitar in white and by saxophone in black. Logarithmic scale for the vertical (frequency) axis is used in (b) and (d) to show low frequencies dominated by bass. The scale of the vertical axis denotes the number of the frequency bin.

3. MONTE-CARLO METHODS

Monte Carlo methods have been widely used in statistical inference and simulations [7]. The main difference to methods listed above is in the *probabilistic* rather than *deterministic* approach. In the methods discussed above, once the initial pattern is fixed, filtering produces the same pattern every time the algorithm is applied. In MC methods, algorithms contain an element of chance, so in a single step the filtering is neither perfect nor reproducible. However, as the algorithm is applied many times, the pattern converges to a steady state. The best analogy can be drawn from minimization of a function by deterministic methods (eg. a method of steepest descend) and by stochastic methods (eg. simulated annealing) [7]. The first method works if there is only a single global minimum, but fails if the function has local minima. The second method is usually time consuming but can identify a global minimum because it searches widely. Many MC methods are based upon analogies with biological systems, particularly genetic models, and on physical systems. In this paper we present two Monte Carlo methods, both using the Metropolis algorithm [8] for calculations.

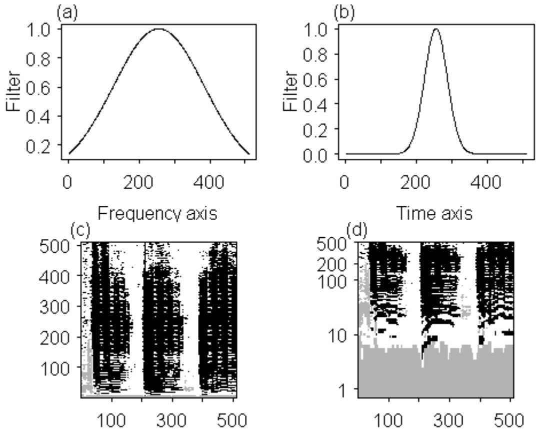

Figure 5. Filter based on two-dimensional Fourier transform using broader filter in the vertical (frequency) axis (a) and narrower filter in the horizontal (time) axis (b). More details are conserved in the vertical (frequency) axis. (c) and (d) show the resulting occupancy map. Regions dominated by bass are shown in grey, by guitar in white and by saxophone in black. Logarithmic scale for the vertical (frequency) axis is used in (d) to show low frequencies dominated by bass.

3.1. Minimization of the boundary

The Local Maximum Rule leads to creation of a pattern (the *occupancy map*) consisting of patches dominated by each signal (figure 1d). The patches often have very irregular boundaries. The goal of smoothing is to convert the pattern in such a way that small patches disappear entirely and the borders of large patterns are smooth. The analogy can be drawn with a set of fluid droplets dropped on a flat surface. The initial pattern might look very patchy, but because of the surface tension, small droplets will join large droplets and irregular droplets will become regular ones with smooth boundaries.

The method that we present here is based upon the concept of minimalization of the boundary length. Consider first a vertical boundary of a patch with a single pixel attached (figure 7a). Consider also an alternative shape, where the single additional pixel is removed (figure 7b). The length of the boundary is clearly larger in the first case, and therefore the removal of the pixel leads to a smaller total length of the boundary. Removal of an isolated pixel leads to an even larger reduction of the total length of the boundary (figure 7c and 7d). This can be used to define a cost function that the MC procedure is minimizing.

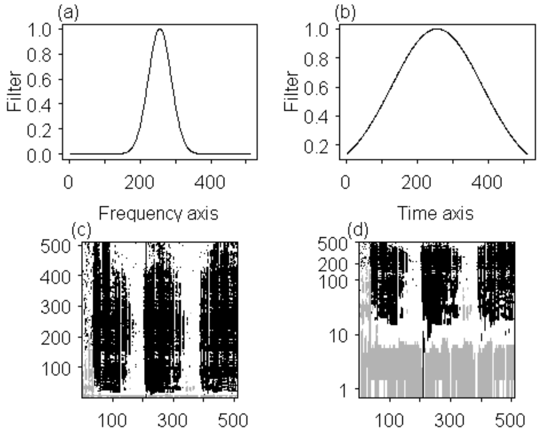

Figure 6. Filter based on two-dimensional Fourier transform using narrower filter in the vertical (frequency) axis (a) and broader filter in the horizontal (time) axis (b). More details are conserved in the horizontal (time) axis. (c) and (d) show the resulting occupancy map. Regions dominated by bass are shown in grey, by guitar in white and by saxophone in black. Logarithmic scale for the vertical (frequency) axis is used in (d) to show low frequencies dominated by bass.

The algorithm proceeds as follows:

1. Select a pixel at random;
2. Change the value of the pixel into another value;
3. Compute the change in the length of the boundary;
4. The change will be accepted with the probability

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

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 regulates the rate at which pixels are removed. The length of updating (the number of times points 1-4 are repeated) is also a parameter: If we stop the simulation early, many isolated points will still be present in the signal. For long simulations, the pattern will be dominated by large patches and the detail will be lost.

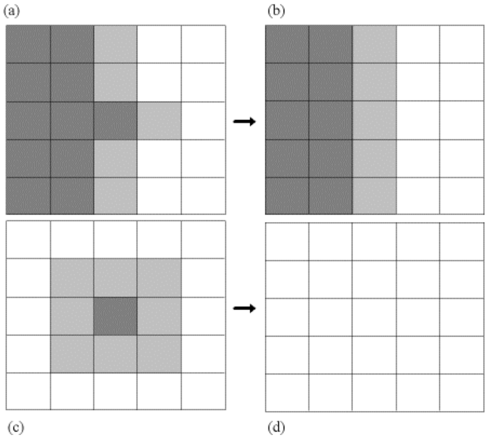

Figure 7. Minimization of the boundary length. Changing a central pixel from dark grey to white reduces the length of the boundary (marked in grey) both in case of a vertical boundary between two areas dominated by different instruments ((a) and (b)) and for a single isolated pixel ((c) and (d)).

However, the procedure is very demanding in processor time and therefore we only show results for 128ms of the signal and limit the frequency range to the lowest 256 bins. Figure 8 shows a comparison for 128ms of the signal chosen near the beginning of the full signal.

3.2. Energy exchange

All methods listed above are essentially based upon the LMR method. The first step in all of them consists of a transformation that replaces in each pixel the information about all instruments with a single instrument. This clearly leads to a loss of information. We therefore need a method that keeps track of all energies.

In this paper we use a simple algorithm that allows efficient filtering by iterative combination and updating of energies for different signals. Biological analogy of the algorithm can be found in species migration and competition. Thus, imagine that each instrument is represented by a different species. Each pixel contains individuals of different species (representing different tracks), and the relative density of each population is given by track energies.

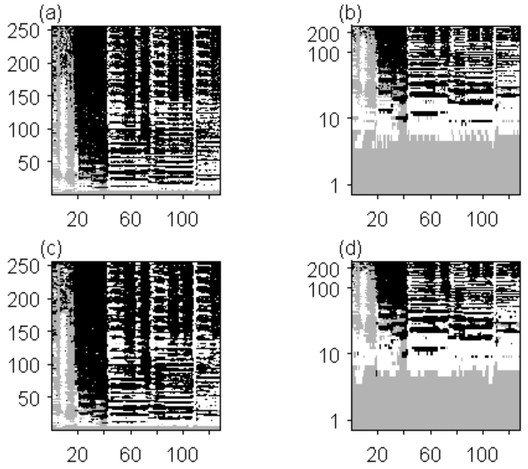

Figure 8. Minimization of the boundary length. Original map is shown in (a) and (b), whereas the filtered map in (c) and (d). Regions dominated by bass are shown in grey, by guitar in white and by saxophone in black. Logarithmic scale for the vertical (frequency) axis is used in (b) and (d) to show low frequencies dominated by bass.

We assume that all species are present at the site, but only one species is dominating – the one with the highest energy. The dominating species can colonize the neighboring pixels. Colonization occurs, providing the dominating species have a higher density than the dominating species in the neighboring pixel. Relative fitness of species in each pixel is determined by a difference in spectral energies. When species in one pixel colonize the neighboring pixel, they will carry with them their energy.

Initially we start with an unfiltered (original) distribution of dominant species $W_{ij} \in \{1,2,3\}$ and corresponding energies $\left\{E_{ij}^{(1)}, E_{ij}^{(2)}, E_{ij}^{(3)}\right\}$. Each step of the analysis consists of the following operations:

1. Choose a pixel (i, j) to be the central pixel;

2. In the neighborhood of the pixel $\{(i-1,j),(i+1,j),(i,j-1),(i,j+1)\}$, find average energies for those tracks that are dominating in the neighboring pixels);
3. Choose the dominating species that have the highest average energy in the neighborhood – this species will then attempt to invade the central pixel;
4. Colonization will not occur if the energy of the invading species is lower than the energy of the species in the central pixel;
5. If the energy of the invading species is larger than the energy of the species in the central pixel, the colonization will occur with the probability given by:

$$1-\exp\left(-a\left(\frac{1}{4}\sum_{k,l\in \text{Neighb}} E_{kl}^{\text{invading}} - E_{ij}^{\text{resident}}\right)\right) \qquad (2)$$

6. The invading species is then colonizing the pixel (i,j), and the new energy of this species at the central pixel is equal to the average energy in the neighborhood.
7. The procedure is then repeated for all pixels;

Points 1-7 are then repeated a number of times. This allows formation of larger clusters of species, because the energy is being updated at each step. After a series of steps no new updates are made and the system reaches a stable state. However, the procedure is very demanding in processor time and therefore we only show results for 128ms of the signal and limit the frequency range to the lowest 256 bins. Figure 9 compares the original signal and the filtered signal.

The smoothing procedure can be changed by adjusting the neighborhood structure and by changing the parameter that regulates the probability of invasion a. The length of updating (the number of times points 1-7 are repeated) is also a parameter: If we stop the simulation early, many isolated points will still be present in the signal.

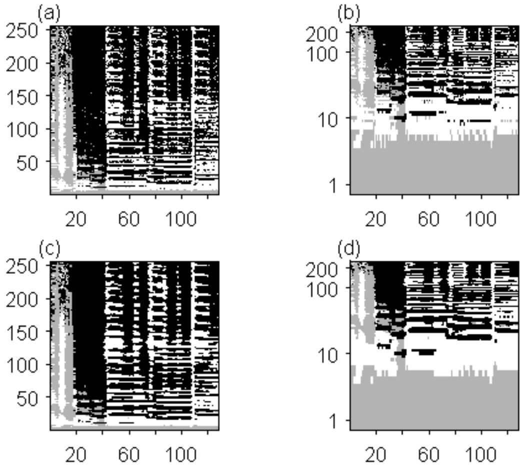

Figure 9. Migrating species method. Original map is shown in (a) and (b), whereas the filtered map in (c) and (d). Regions dominated by bass are shown in grey, by guitar in white and by saxophone in black. Logarithmic scale for the vertical (frequency) axis is used in (b) and (d) to show low frequencies dominated by bass.

4. CONCLUSIONS

Each of the presented methods of determining the areas in the time-frequency plane for the Selective Mixing of Sounds has its specific parameters allowing for adjustment for best perceptual results.

In [1] a specific “asymmetric” Moore neighborhood of variable order along the frequency scale was used. The characteristic effects of selective mixing on sound, summarized in [1] are preserved in all of the methods presented here, according to the informal listening tests. The perceptual differences between the methods are subtle and the proper perceptual evaluation of these methods require numerous further listening tests.

5. REFERENCES

[1] P. Kleczkowski, “Selective Mixing of Sounds”, 119th Convention of the Audio Engineering Society, New York, October 2005, preprint 6552.

[2] A.S. Bregman, *Auditory Scene Analysis,* MIT Press, Cambridge, 1990.

[3] Kelly M.C., Tew A.I., "The continuity illusion in virtual auditory space", 112th Convention of the Audio Engineering Society, Munich, May 2002, preprint 5548.

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

[5] Kelly M.C., Tew A.I., "The significance of spectral overlap in multiple-source localization", 114th Convention of the Audio Engineering Society, Amsterdam, March 2003, preprint 5725.

[6] Bloomfield, P. *Fourier Analysis of Time Series: An Introduction*. Wiley, New York, 1976.

[7] Otten, R.H.J.M. and van Ginneken, L.P.P.P., *The Annealing Algorithm*, Kluwer, Boston, 1989.

[8] Metropolis, N., Rosenbluth, M., Teller, A., and Teller, E. "Equation of State Calculations by Fast Computing Machines", Journal of Chemical Physics, June 1953, vol. 21, pp. 1087-1092