ARCHIVES OF ACOUSTICS
33, 4 (Supplement), 91–98 (2008)

SELECTIVE MIXING OF A SYMPHONIC ORCHESTRA RECORDING

Piotr KLECZKOWSKI

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

(*received June 15, 2008; accepted October 24, 2008*)

The selective method of mixing of sounds, developed by the author, has been applied to a multitrack recording of a symphonic orchestra. This method, still under development, offers a slight increase in the perceived clarity of recordings when compared to normal mixing. Fifteen stereophonic tracks of a symphonic orchestra recording have been processed. At first, the algorithm used earlier by the author for the recordings consisting of five and six tracks was used. After considering the remarks of listeners of informal listening tests, two modifications of that algorithm, specific for the recordings of a symphonic orchestra, have been introduced. Both modifications consisted in specific processing of selected groups of the tracks. During the final listening test, 32 untrained listeners compared the normal mix with the basic algorithm and with the two advanced algorithms of selective mixing. The advanced algorithms worked better than the basic one, indicating the direction for further improvement.

Keywords: audio signal processing, psychoacoustics.

1. Introduction

1.1. The psychoacoustical foundations

The human ear usually receives mixtures of sounds overlapping in the time-frequency plane. The phenomenon of masking, mostly investigated with the use of laboratory test signals like tones or noises, does not fully account for the human reception of sounds encountered in real life, including music. The author had performed experiments, simulating the raising of the threshold of masking where the masker and the maskees were sounds of musical instruments. Substantial amounts of artificial increase of masking did not affect the perceived quality of mixes of sounds processed in this way. Continuing that research, the author processed the mix of two musical instruments in the time-frequency domain and found that in each area of the time-frequency plane the elimination (muting) of the weaker element of two sounds has negligible effect on the perceived sound of either of the two instruments [11, 13]. The similar phenomenon holds when the number of instruments is increased. The author experimented with mixes of 3, 4, 5 and 6 instruments. The following conclusion was given in [11]: “in any small area of the time-frequency plane all respective segments of sounds can be removed except the sound with the highest energy in that area, and the quality of sound remains satisfactory”. However, as the number of instruments increases, the audible difference between the original mix and the processed mix becomes gradually more perceivable. On the one hand, more details in the sounds are heard, but on the other hand the mix becomes a bit “dry” and slight artefacts are occasionally perceived [11].

1.2. Related research

BREGMAN in [4] gives an extensive overview of a family of psychoacoustic effects called together the “continuity illusion” (“auditory continuity” by some other researchers). In his words “...the experimenter deletes parts of a softer signal and replaces them with a louder sound, and the listener hears the softer sound as continuing unbroken behind the louder one”. WARREN [14] considers the continuity illusion as a mechanism compensating for masking. There is likely a relation between the effect described in Sec. 1.1 and the illusion of continuity, but not all the rules given by Bregman and Warren as conditions for continuity illusion to occur were satisfied in the experiments mentioned in Sec. 1.1.

Two teams of researchers investigated the possible applications of an operation of muting or removing parts of sounds in the areas of spectro-temporal overlap. BAUMGARTE and FALLER [1, 2, 5, 6] proposed a technique called Binaural Cue Coding, aimed at coding of multichannel audio in one channel plus side information channel. The authors argue that their BCC scheme has advantages over low bit rate stereo and multichannel compression. An important element of their technology consists in neglecting spectro-temporal elements with lower energy. KELLY and TEW [7–9] performed an experiment similar to this authors’ mentioned in Sec. 1.1, using two signals: a vocal track and an accompanying piano track, and compared them in separate regions of the time-frequency plane. Their finding was that it was possible to remove the weaker signal in a region only if its level was lower by at least 15 dB, should no perceptible degradation of the mix occur. This is in contrast with the findings of this author mentioned above. In [8] they also investigated the effect of this same operation on localization accuracy and found that it was possible to remove the weaker signal in all regions without the effect on localization. In [9] they proposed, like BAUMGARTE and FALLER, to use their findings to increase the efficiency of coding multichannel audio.

1.3. The purpose of this research

In [11] it was shown that by a refinement of a simple, elimination based processing according to Sec. 1.1, a specific algorithm of mixing is obtained, producing mixes of 5 and 6 instruments preferred by many listeners. This method of mixing has been called the selective mixing of sounds. There were two aims of the research reported here. First, the investigation of the perceptual effect of spectral elimination when the number of instruments is increased. Second, the development of a specific technique for the selective mixing appropriate for mixing more than just a couple of instruments.

2. The implementation of selective mixing for a low number of audio tracks

The basic block diagram of the procedure of selective mixing is shown in Fig. 1.

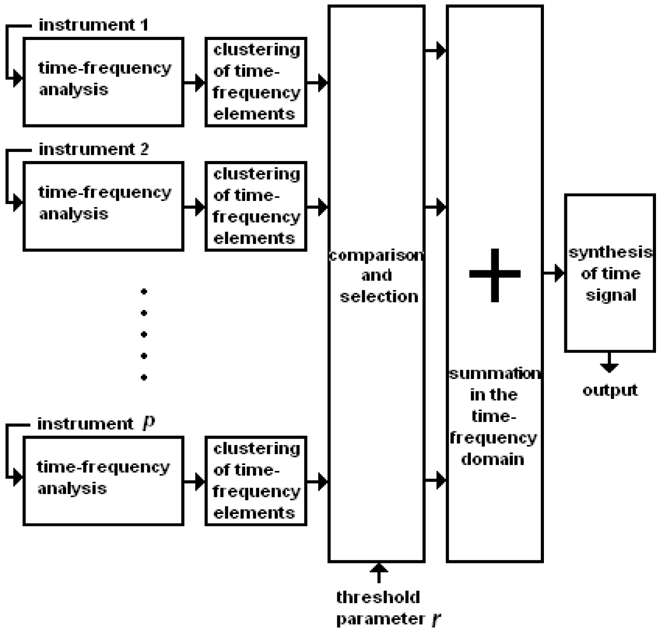

Fig. 1. The block diagram of the basic implementation of selective mixing.

The input signals are the separate tracks. Their relative levels should be set so that their sounds are balanced in the normal mix and this must be performed by a human operator (mixing engineer). The equalization and the control of dynamics can be adjusted as during normal mixing.

All the input signals are time-frequency analyzed. The author has used the algorithm presented in [10] but other methods of time-frequency analyses can be used. The essential process of selection in its most basic form is performed individually in each time-frequency cell, according to the formula:

$$|F|_{k,n,m} = \max\left\{|F|_{k,n,1}, |F|_{k,n,2}, \dots |F|_{k,n,p}\right\}, \tag{1}$$

where F – time-frequency distribution coefficient, k – frequency bin index of a time-frequency cell, n – time frame index of a time-frequency cell, m – instrument (audio track) index, p – no. of instruments.

The basic selection according to (1) should not be used, as rapid switching between many individual cells belonging to different instruments occur, causing nonlinear distortion. Rather, larger time-frequency areas with smoother boundaries are recommended.

Various strategies of clustering individual cells into larger areas can be used. Some of them were proposed in [12]. A high-order Moore neighbourhood rule was used in [11] and in this work, according to:

$$M_{k,n,m} = \sqrt{\frac{\sum\limits_{k=k-i}^{k=k+i}\sum\limits_{n=n-j}^{n=n+j} F^2_{k,n,m}}{(2i+1)(2j+1)}}, \tag{2}$$

where M – average amplitude of time-frequency coefficient in a neighbourhood, k – frequency bin index of a centre of a neighbourhood, n – time frame index of a centre of a neighbourhood, m – instrument (audio track) index.

The values of i and j depend on the number of a frequency bin, so that the widths of the areas along the frequency scale depend on frequency, thus approximating the widths of the critical bands.

This leads to the modified form of the selection process in (1), where $\max\{M_{k,n,1}, M_{k,n,2}, ..., M_{k,n,p}\}$ is used to determine the actual value of the instrument index m, i.e. to determine the instrument with the highest value of average energy in a neighbourhood centred at the k, n point in the time-frequency plane. The Moore neighbourhood averaging of (2) increases the local areas of dominance in the time-frequency plane, reducing considerably the number of points of switching between the instruments. The basic selection process is given by the formula:

$$F_{k,n,m} = 0, \quad \text{for} \quad M_{k,n,m} \neq \max\{M_{k,n,1}, M_{k,n,2}, ..., M_{k,n,p}\} \tag{3}$$

In order to soften the effect of selective mixing, i.e. to eliminate occasional artefacts, the above formula should be modified, so that the instruments which have considerable (although not the highest) energy in a particular area of the time frequency plane are not eliminated. This is accomplished by using the following rule:

$$F_{k,n,m} = 0, \quad \text{for} \quad M_{k,n,m} < r \cdot \max\{M_{k,n,1}, M_{k,n,2}, ..., M_{k,n,p}\} \tag{4}$$

where r – is a threshold, determining the relative level of elimination of instruments in the time-frequency plane, $0 < r < 1$.

3. Selective mixing of a symphony orchestra recording

3.1. The application of the basic algorithm

The input material for the experiment was a multitrack recording of a symphony by Mozart, consisting of the following, mostly stereophonic tracks: 4 tracks of the violins, 2 tracks of the violas, 2 tracks of the cellos, and single tracks of the flute, the oboe, the corn, the bassoon and the double bass. In addition, there was a main XY track and an ambience track. The relative levels of all the tracks have been set by a professional mixing engineer. In the selective mixing of stereo tracks the procedure of Fig. 1 is applied independently to all the left channels and to all the right channels. The XY and ambience tracks required special handling, since they contained sounds of all instruments, while the process of selective mixing assumes the "competition" of different instruments. Two alternative techniques are possible here:

1. Both tracks are added to the recording resulting from the selective process.
2. Both tracks are not present in the output recording and the artificial reverberation is used at the output of the selective mixing process.

The choice of the option does not seem to be relevant to the selective mixing, according to informal listening tests. Option 1 has been used in the listening tests as more consistent with the original mix.

The listening test consisted in comparing the sound of the normal mix and the selective mix, according to the scheme of Fig. 1. The selection according to (3) was used. It was assumed that the function of softening of the selection process according to (4) was performed by the XY and ambience tracks, which contained full sounds.

Six trained listeners participated in an informal experiment. The result was negative for the selective mixing, as it was preferred only by one of the listeners, while all the others preferred the normal mix. After the discussion with the listeners it was concluded, that besides occasional artifacts, switching occurs between the instruments playing unison, especially between those belonging to the same group. The analysis of appropriate plots revealed, that the switching took place at intervals between 100 and 200 msec.

3.2. The use of sub-mixes as input signals

In order to avoid the effect of switching it was necessary to eliminate the mutual competition between the instruments of similar time-frequency distributions. The block diagram of the modified procedure is shown in Fig. 2.

The simple approach consists in performing sub-mixing of instruments in the appropriate groups, so that those groups are used as individual input signals.

The analysis of melody lines and the spectral distrubutions indicated the following two groups. One contained all four violin tracks plus two viola tracks, and the other contained two cello tracks.

3.3. Intra-group competition and dynamic control of the threshold of elimination

Another strategy is presented in the form of a block diagram in Fig. 3.

The division to sub-groups has been maintained there, but an intra-group competition has been introduced. This competition was made considerably softer that in the basic algorithm of Fig. 1, by using a low value of the threshold parameter r, with the purpose of obtaining mixes rather than individual instruments in large areas of the time-frequency plane, in order to avoid frequent switching. After the intra-group selection the time-frequency signals of individual instruments are summed in the frequency domain, and their sum competes as one group with the rest of individual instruments. Thus the balance between all instruments, similar to that of the basic algorithm is maintained, while the mutual competition within the groups has been reduced.

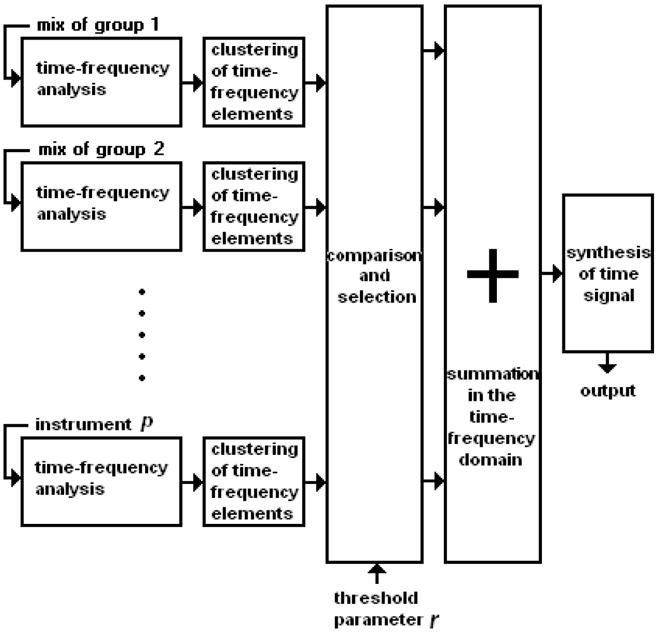

Fig. 2. The block diagram showing the procedure with sub-mixes at the input.

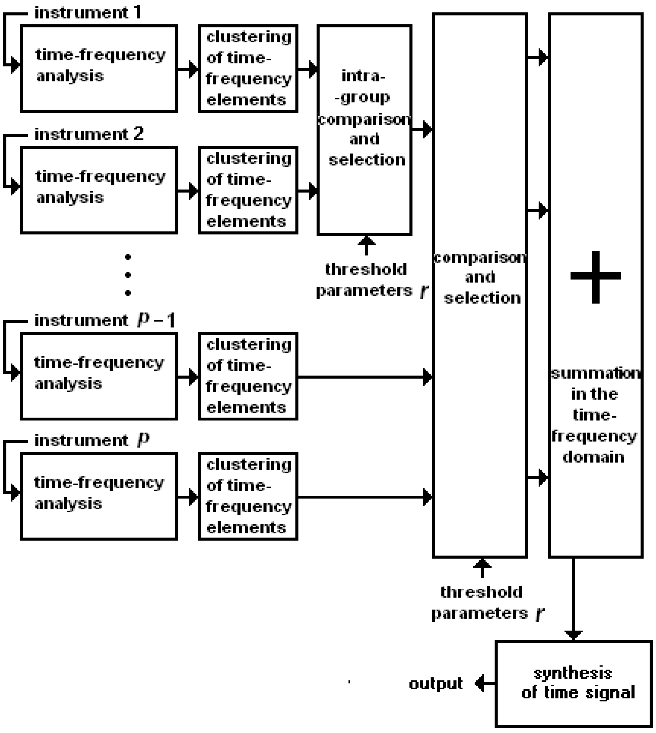

Fig. 3. The block diagram including intra-group competition and the dynamic control of parameter r.

One more control tool is in making the threshold r time dependent, so that r can be lowered at the moments when artifacts or switching occurs. The control of r, both fixed in time and dynamically changed must be performed by a human operator, just like other operations in the process of mixing. The same grouping as in Sec. 3.2. was used for controlling r. There were two short periods (of about 1.5 sec. and about 0.5 sec. respectively) when r was lowered from its constant value. The actual values of r have been subjectively set by the author, in order to achieve good overall sound, so they are given just as an example. The values of r are given in Table 1.

Table 1. Values of the threshold r used in the test example.

instrument or group	violins + violas	cellos	flute	oboe	corn	bassoon	double bass
value of r	0.5 in-group	0.5 in-group	0.8	0.7	0.8	0.8	0.9
	0.75 0.3 lowered	0.7 0.3 lowered					

3.4. Listening tests

The listening tests consisted in the comparison of the normal mix with the basic selective mix and both versions of the advanced selective mix, of Sec. 3.2 and 3.3 respectively. The pair: the original mix and one type of the selective mix was compared in each individual experiment. Both the order of items in pairs and the order of the three experiments were randomised. Thirty two untrained listeners participated in the experiment, divided into three groups of 10–12 participants. The test examples were reproduced by a pair of Genelec 1038A monitors. A modified "paired rating hidden reference" test paradigm was used [3]. Since the perceived difference between the original and the selective versions was subtle, instead of a 7-point category rating scale the listeners were just asked about their preferred version. The answer "I am not sure" was allowed. The participants were informed that the two versions differed only by the mixing technique. The results are shown in Table 2. The result of the comparison "original mix – basic selective mix" has a high significance level ($\alpha = 0.001$, computed according to the binomial distribution), but the significance levels for the other two comparisons were inadequate to draw meaningful conclusions from the test. It can be concluded that there was no perceptual difference with the "3.2" algorithm, while some difference was observed with the "3.3" algorithm, with balanced priorities.

Table 2. Results of the listening test.

no. of participants	no. of not sure	no. choosing original mix	no. choosing basic selective mix	no. choosing selective mix according to 3.2	no. choosing selective mix according to 3.3
32	3	26	3	–	–
32	17	9	–	6	–
32	6	12	–	–	14

4. Conclusions

The psychoacoustical effect, where the removal of large parts of musical tracks in the time-frequency domain may be not perceived in the mix at all, has previously been confirmed with the examples of up to 6 audio tracks. This research showed that this number can be increased to at least 13 tracks. However, simple algorithm of selective mixing (of Fig. 1), previously found to be able to improve some aspects of the sound for up to 6 tracks, brings sound degradation when used with 13 tracks. At this amount of tracks, more sophisticated techniques of selection should be used. Two techniques have been proposed, and it was shown that they were able to eliminate degradation, but no statistically significant improvement of sound has been obtained yet.

References

[1] Baumgarte F., Faller C., *Why Binaural Cue Coding is better than intensity stereo*, 112th Audio Eng. Soc. Conv, Munich, May 2002.

[2] Baumgarte F., Faller C., *Design and Evaluation of Binaural Cue Coding Schemes*, 113th Audio Eng. Soc. Conv, Los Angeles, Preprint 5706, October 2002.

[3] Bech S., Zacharov N., *Perceptual Audio Evaluation*, Wiley 2006.

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

[5] Faller C., Baumgarte F., *Binaural Cue Coding applied to stereo and multi-channel audio compression*, 112th Audio Eng. Soc. Conv, Munich, Preprint 5574, May 2002.

[6] Faller C., Baumgarte F., *Binaural Cue Coding applied to audio compression with flexible rendering*, 113th Audio Eng. Soc. Conv, Los Angeles, Preprint 5686, October 2002.

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

[8] 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, pp. 9–12, Nov. 2002.

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

[10] Kleczkowski P., *Acoustic Signal Expansion in Multiple Trigonometric Bases*, Acustica/Acta Acustica, **88**, 526–535 (2002).

[11] Kleczkowski P., *Selective Mixing of Sounds*, 119th Convention of the Audio Engineering Society, New York, Preprint 6552, October 2005.

[12] Kleczkowski P., Kleczkowski A., *Advanced Methods for Shaping Time-Frequency Areas for the Selective Mixing of Sounds*, 120th Convention of the Audio Engineering Society, Paris, Preprint 6718, May 2006.

[13] Kleczkowski P., *Some experiments on extreme masking*, Archives of Acoustics, **31**, 4 (Supplement), 409–416 (2006).

[14] Warren R.M., *Perceptual restoration of obliterated sounds*, Psychol. Bull., **96**, 371–383 (1984).