Prioritizing signals for selective real-time audio processing

Emmanuel Gallo, Guillaume Lemaitre, Nicolas Tsingos

PRIORITIZING SIGNALS FOR SELECTIVE REAL-TIME AUDIO PROCESSING

Emmanuel Gallo[1,2]*, Guillaume Lemaitre*[1] *and Nicolas Tsingos*[1]

[1]REVES-INRIA and [2] CSTB,
Sophia Antipolis, France.
Emmanuel.Gallo@sophia.inria.fr

ABSTRACT

This paper studies various priority metrics that can be used to progressively select sub-parts of a number of audio signals for real-time processing. In particular, five level-related metrics were examined: RMS level, A-weighted level, Zwicker and Moore loudness models and a masking threshold-based model. We conducted a pilot subjective evaluation study aimed at evaluating which metric would perform best at reconstructing mixtures of various types (speech, ambient and music) using only a budget amount of original audio data. Our results suggest that A-weighting performs the worst while results obtained with loudness metrics appear to depend on the type of signals. RMS level offers a good compromise for all cases. Our results also show that significant sub-parts of the original audio data can be omitted in most cases, without noticeable degradation in the generated mixtures, which validates the usability of our selective processing approach for real-time applications. In this context, we successfully implemented a prototype 3D audio rendering pipeline using our selective approach.

1. INTRODUCTION

Many applications ranging from video games to virtual reality or visualization/sonification require processing large number of audio signals in real-time. For instance, modern video games must render a large number of 3D sound sources using some form of spatial audio processing. Furthermore, each source's audio signal may itself be generated as a mixture of a number of sub-signals (e.g., a car-noise is a composite of engine and tire/surface noise) driven from real-time simulated physical parameters. The number of audio signals to process may often exceed hardware capabilities. Priority schemes that select the sounds to process according to a preset importance value are a common way of using hardware more efficiently, for instance by managing the limited number of hardware channels on a dedicated sound card. Usually, this value is determined by the sound designer at production time and might further be modulated by additional effects at run-time, such as attenuation of the sound due to distance or occlusion.

This paper is focused on the problem of automatically prioritizing audio signals according to an importance metric, in order to selectively process these signals. Such a metric can then be used to tune the processing "bit-rate" in order to fit a given computational budget: for instance, allocating a budget number of arithmetic operations to a complex signal processing task (e.g., a combination of mixing, filtering, etc.) involving a large number of source signals.

Figure 1 shows a basic example application where four speech signals have been prioritized according to a loudness metric and a mix has been generated simply by playing back the single most-important signal per processing frame (highlighted in yellow). Such an approach could typically be used for hardware voice management in video games.

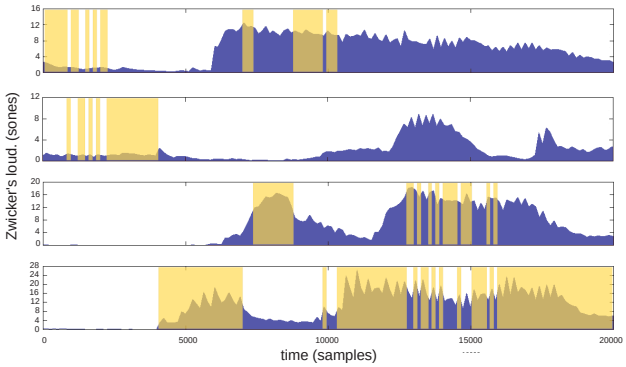

Figure 1: Four speech signals prioritized according to a loudness metric computed over successive short time-frames. The single most important frame across time is highlighted in yellow.

This paper presents a comparative study of several metrics that can be used to prioritize signals for selective real-time processing of audio signals. In section 2, we start by reviewing previous work related to scalable and progressive audio processing. A coarse-grain selective processing algorithm is described in section 3. In particular, several metrics that can be used to prioritize the audio signals and selectively allocate the required operations are discussed in section 3.1. Our selective processing algorithm is demonstrated in the context of a time-domain pipeline comprising mixing and simple filtering operations in section 3.2. Results of a pilot subjective study are presented in section 4 that support the applicability of our technique. We finally discuss our approach and outline other possible applications of our prioritization scheme before concluding.

2. RELATED WORK

While parametric, progressive and scalable codecs are a key research topic in the audio coding community [1, 2, 3], few attempts to date have been made to design scalable or selective approaches for real-time signal processing.

Fouad et al. [4] propose a level-of-detail rendering approach for spatialized audio where the sound samples are progressively generated based on a perceptual metric in order to respect a budgeted computing time. When it is elapsed, missing samples are interpolated from the calculated ones. As they prioritize signals

according to their overall energy, such a scheme will fail at capturing large energy variations through time within the signal itself.

Wand and Straßer [5] proposed a multi-resolution approach to 3D audio rendering. At each frame of their simulation, they use an importance sampling strategy to randomly select a sub-set of all sound sources to render. However, their importance sampling strategy also does not account for the variations in signal intensity. Such effects might be much more significant (factors of 10 or more can be easily observed on speech signals for instance) than variations in the control parameters such as distance attenuation, etc. since the latter usually vary smoothly and slowly through time (except for very near-field sources).

In a previous work [6], we proposed a framework for 3D audio rendering of complex virtual environments in which sound sources are first sorted by an *importance* metric, in our case the *loudness level* of the sound signals. We use pre-computed descriptors of the input audio signals (e.g., energy in several frequency bands through time) to efficiently re-evaluate the importance of each sound source according to its location relative to the listener. Hence, loudness variations within the signals are properly accounted for. The priority metric was used to determine inaudible sources in the environment due to auditory masking and group sound sources to optimize spatialization. This paper extends this approach by comparing several priority metrics and their subjective effect on selective processing of audio signals even for cases where removed sub-parts of the signals are above masking threshold.

Other scalable approaches based, for instance, on modal synthesis, have also been proposed for real-time rendering of multiple contact sounds in virtual environments [7, 8, 9]. Similar parametric audio representations [1, 2] also allow for scalable audio processing (e.g. pitch shifting or time-stretching, frequency content alteration, etc.) at limited additional processing cost, since only a limited number of parameters are processed rather than the full Pulse Code Modulation (PCM) audio data. However, this approach might imply real-time coding and decoding of the sound representations. As parametric representations are not widely standardized and commonly used in interactive applications, available standard hardware decoders do not usually give access to the coded representation in a convenient form for the user to further manipulate. Eventhough processing in coded domain might be achieved through modified software implementation of standard audio codecs (e.g. MPEG-1 layer 3, MPEG-2 AAC) [10], the overhead due to partial decoding would probably be overwhelming for a real-time application handling many signals.

3. SELECTIVE AUDIO PROCESSING

We propose a coarse-grain selective audio processing framework that can be separated into two steps : 1) we assign a priority to each frame of the input signals and 2) we select the frames to process by decreasing priority order until our pre-specified budget is reached. Remaining frames are simply discarded from the final result. Both steps are applied at each processing frame to produce a frame of processed output signal. The following sections detail both steps.

3.1. Priority metrics

In our approach, as well as others we described in section 2, processing management is driven by a given importance metric. The choice of this metric is then a crucial step: the audibility of the artefacts introduced by any processing optimizations will depend on its quality.

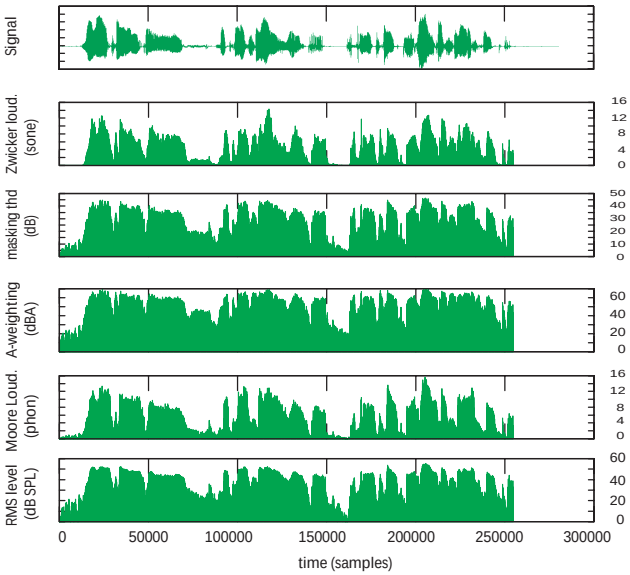

Figure 2: Several priority metrics calculated for an example speech signal using 3 ms-long frames.

Loudness seems a good candidate since it has been shown to be closely related to masking phenomena [11, 12]. Using loudness as an importance metric might hence allow important maskers to be processed first. But one can imagine that weighting may be more efficiently performed on the basis of more cognitive aspects. For instance, in the context of a collision avoidance experimental setup, Robert Graham [13] noticed that faster braking reaction times were measured when drivers were warned by car horns sounds, even if they were less loud than other tested sounds. There is a vast literature aiming at building psychoacoustic relationships between acoustic parameters of a sound and its so-called *urgency* (see Stanton and Edworthy [14] and [15] for an overview). Derivations of these urgency metrics may form a more cognitive-founded importance metric.

As a starting point, this paper examines the ability of several level-related metrics to optimize audio processing. In particular, we evaluated the following importance metrics:

1. RMS level, expressed in dB SPL,
2. *A*-weighted level, expressed in dBA [16],
3. Moore, Glasberg & Baer's loudness level [17], expressed in *phons*, calculated assuming a stimulus is a band-limited noise.
4. Zwicker's loudness [18], expressed in *sones*,
5. "Masking level" model defined as the level of the source minus a masking-threshold offset, expressed in relative dB (a masking threshold of -3 dB indicates that sounds with a energy weaker than half the energy of the masking sound will be masked), predicted from the *tonality* index of the signal [19, 20]. Tonality index is typically derived from a *spectral flatness measure* and indicates the tonal or noisy nature of the signal [21].

Each metric is evaluated for short processing frames along our test signals, typically every 3 to 23 ms (i.e., 128 to 1024 samples at 44.1kHz). Results were not significantly different for the various frame sizes. Smaller frames give better time-resolution and can result in more optimal interleaving of the signals during the processing step. However, frames too short can result in highly degraded audio information since interleaved signals will no longer be recognizable, a problem closely related to the illusion of continuity [22]. Figure 2 shows a comparison of several loudness metrics evaluated on a fragment of speech signal.

Table 1 shows the average rank correlation obtained with various metrics on three different mixtures of speech, ambient and music signals. Rank correlation measures how correlated the orderings obtained with the various metrics are. As can be seen in this table, results appear to be dependent on the type of signals. For speech and ambient sounds, metrics are correlated although not strongly. For the musical mixture, results are more pronounced showing stronger correlation between Zwicker's and Moore's loudness models and very low correlation between loudness models and all the others.

speech	Zwicker loud.	mask. thr.	Moore loud.	RMS level	A-weight.
Zwicker loud.	1 (0)	0.37 (0.22)	0.57 (0.29)	0.40 (0.22)	0.35(0.23)
mask thr.	0.37 (0.22)	1 (0)	0.54 (0.22)	0.73 (0.22)	0.56 (0.31)
Moore loud.	0.57 (0.29)	0.54 (0.22)	1 (0)	0.54 (0.23)	0.36 (0.21)
RMS level	0.40 (0.23)	0.73 (0.22)	0.54 (0.23)	1 (0)	0.54 (0.31)
A-weight.	0.35 (0.23)	0.56 (0.31)	0.36 (0.21)	0.54 (0.31)	1 (0)
ambient	Zwicker loud.	mask. thr.	Moore loud.	RMS level	A-weight.
Zwicker loud.	1 (0)	0.40 (0.18)	0.44 (0.18)	0.42 (0.19)	0.37 (0.17)
mask thr.	0.40 (0.18)	1 (0)	0.48 (0.17)	0.51 (0.18)	0.35 (0.18)
Moore loud.	0.44 (0.18)	0.48 (0.17)	1 (0)	0.47 (0.17)	0.37 (0.17)
RMS level	0.42 (0.19)	0.51 (0.17)	0.47 (0.17)	1 (0)	0.33 (0.18)
A-weight.	0.37 (0.17)	0.35 (0.17)	0.37 (0.17)	0.33 (0.18)	1 (0)
music	Zwicker loud.	mask. thr.	Moore loud.	RMS level	A-weight.
Zwicker loud.	1 (0)	0.05 (0.12)	0.42 (0.12)	0.04 (0.11)	0.03 (0.10)
mask thr.	0.05 (0.11)	1 (0)	0.04 (0.11)	0.42 (0.09)	0.40 (0.10)
Moore loud.	0.42 (0.12)	0.04 (0.11)	1 (0)	0.02 (0.10)	0 (0.10)
RMS level	0.04 (0.11)	0.42 (0.10)	0.02 (0.10)	1 (0)	0.36 (0.10)
A-weight.	0.03 (0.10)	0.40 (0.10)	0 (0.10)	0.36 (0.10)	1 (0)

Table 1: Rank correlation matrices for three test mixtures of speech, ambient and musical signals. Rank correlation was calculated using Spearman's formula [23] and averaged over all frames of the mixture. Its variance across frames is also given in brackets.

3.2. Selective processing algorithm

Our budget allocation algorithm is designed for real-time streaming applications. Hence, it has to be efficient and has to find a solution locally at each processing frame. To do this, the importance of each frame of the signal is evaluated and until our computational budget is reached, the algorithm selects which sub-parts of the signals should be processed, by decreasing priority value, using a greedy approach (i.e, taking the best immediate, or local, solution). An example is shown in Figure 1. The result is thus constructed as an interleaved mixture of the most important frames in all signals. To avoid artefacts during the reconstruction step, an overlap-add method (3ms frames with 10% overlap) was used. Another example is shown in Figure 3. Selected frames for different budgets are highlighted. As can be seen in the figure, our approach directly accounts for any sparseness in the mix by removing input frames below audibility threshold from the final mix. This might already result in a significant gain. For the various mixtures we used (ambient sounds, music and speech), we estimated that 0.7% to 33% of the input frames could be trivially removed (0.7% for ambient sounds, 24.5% for music and 33% for speech).

To improve the frequency resolution of our approach, we can further evaluate the priority metric for a number of sub-bands of the signals. In our experiments, we used four sub-bands corresponding to 0-500 Hz, 500-2000 Hz, 2000-8000 Hz, 8000-22000 Hz and treated each sub-band as if it were an additional input sound signal to prioritize. This would be typically useful for applications performing some kind of sub-band correction of the audio signal (e.g., equalizers). The required band-pass filtering can then be performed only on the selected sub-parts of the signal.

3.3. Integration within a real-time processing framework

Although most of the level-related priority metrics we used cannot be directly evaluated in real-time for large numbers of audio streams, they can be efficiently computed from additional descriptors stored with the audio data, in a manner similar to [6]. Loudness information, in particular, can be retrieved from precomputed loudness tables, energy levels and tonality indices stored with each corresponding frame of input audio data. This information may also be stored for several sub-bands of the signal. Such an approach allows us to further modulate the importance value in real-time depending on various other effects affecting the signal during the simulation. This necessary information is quite compact (typically about 1 to 4 Kb/sec of input audio data) and can be interleaved with standard PCM audio data for streaming or kept resident in memory for random access while the PCM data is streamed on-demand.

Our coarse-grain selective algorithm integrates well within standard time-domain audio processing pipelines. We evaluated it in the context of a 3D audio processing application for virtual reality. In this case, the signals of each virtual sound source undergo filtering and resampling operations to simulate propagation effects (atmospheric scattering, occlusion, Doppler shift, etc.) and binaural hearing (e.g., HRTF filtering) before being combined to produce the final mix. We implemented a scalable 3D audio processing pipeline implementing these effects using time-domain resampling and attenuation over several sub-bands of each source signal, computed using second-order biquad filters. Using our selective pipeline, we were able to process the signals using a budget number of operations resulting in a computing speed-up directly proportional to the selected budget. As sources of decreasing priority are processed, a complementary solution is to simplify the operations (for instance, using linear resampling instead of better quality spline-based resampling) rather than maintain high-quality processing for all selected frames and simply drop low priority frames. Example movie files demonstrating the approach are available at:
http://www-sop.inria.fr/reves/projects/scalableAudio/.

4. PILOT SUBJECTIVE EVALUATION

In order to evaluate subjective differences between the various metrics we ran a pilot evaluation study described in the following sections.

4.1. Experimental conditions

Subjects: 18 subjects (10 women and 8 men, 19 to 48 years old) volunteered as listeners. All reported normal hearing. Most of them were computer scientists, very few with any experience in acoustics or music practice. None of them was familiar with audio

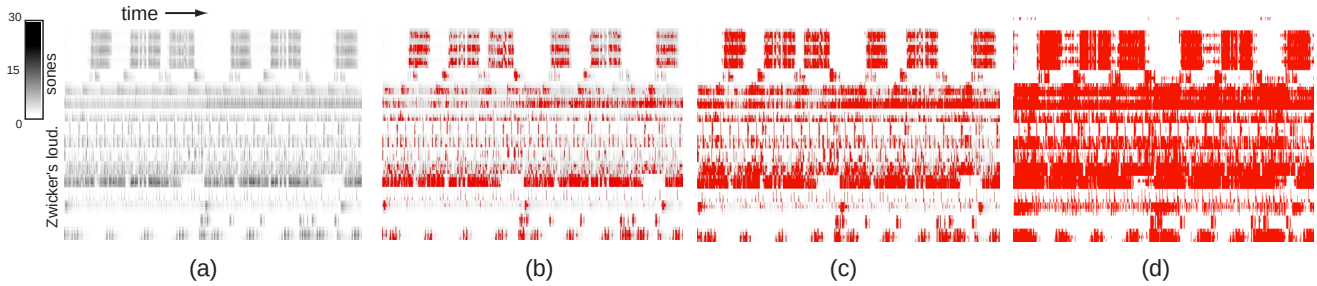

Figure 3: (a) Loudness values (using Zwicker's loudness model) through time for the 17 tracks of a musical mix. Each track was selectively filtered and processed into 4 frequency sub-bands resulting in 68 signals to prioritize. (b) Priority map showing the first 12.5% most important frames highlighted in red. (c) First 25% most important frames and (d) first 50% most important frames.

coding techniques, nor were regular users of mp3 or other coded-audio standards.

Stimuli: Three mixtures of various *types of signals* were generated: 1) a multi-track musical mix, 2) male and female Greek, French and Polish speech and 3) ambient sounds. The mixtures were created respectively from 17, 6 and 4 recordings separated into four sub-bands, resulting in 68, 24 and 16 signals to prioritize. Mixtures were generated at three *resolutions*, selecting the most important frames according to our priority metrics, using only 50%, 25% and 12.5% of the input signal data. Five different priority *metrics* (see section 3.1) were tested. A total of 45 stimuli (3 types of signals * 3 resolutions * 5 metrics) were hence created. All signals were presented at CD quality (44.1 kHz sampling rate and 16 bits quantization)[1].

Apparatus: We ran the test on a laptop computer using an in-house test program (see Figure 4). It was conducted using headphone presentation in a quiet office room. *Sennheiser HD600* headphones were used (diotic listening), calibrated to a reference listening level at eardrum (100 dB SPL). The sounds were stored on the computer hard drive and played through the SigmaTel C-Major integrated sound-board. They were played back at a comfortable level.

Procedure: The subjects were given written instructions explaining the task. They were asked to rate the 45 resulting output mixtures relative to the corresponding reference mix. We used the ITU-R[2] recommended *triple stimulus, double blind with hidden reference* technique, previously used for quality assessment of low bit-rate audio codecs [24]. Subjects were presented with three stimuli, R, A and B, corresponding to the reference, the test stimulus and a hidden reference stimulus (the hidden reference was the reference itself, without any alteration)[3]. Test stimuli were presented to each subject in a different (random) order. The hidden reference was randomly assigned to button A or B. Our test program automatically kept track in an output log file of the presentation order and the marks given respectively to the stimulus and the hidden reference signal for each test. The output of the procedure was then two scores: one for the hidden reference, and one the for the test stimulus. After the ITU-R standard, the judgment value used for further analysis was the difference between the scores of the hidden reference and test stimulus. Hence, a value of zero indicates than no difference was perceived between the reference and the test sound. A positive score indicates that an annoying difference was heard between the test sound and the reference sound, and a negative score indicates that the test sound was better rated than the reference sound.

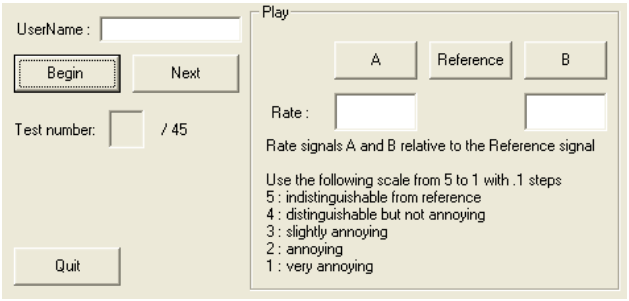

Figure 4: Snapshot of the interface designed for our listening tests.

Subjects could switch between the three stimuli at any time during playback by pressing the corresponding buttons on the interface (see Figure 4). They were asked to rate differences between each test stimuli (A and B) and the Reference from "imperceptible" to "very annoying", using a scale ranging from 5.0 to 1.0 (with one decimal) [25].

After the test, subjects were invited, during a semi-guided interview, to describe the differences that they heard between the processed and the original sounds.

4.2. Analysis

Correlations between the subjects: All subjects raw judgments were significantly correlated ($p < 0.01$) except for one who was removed from further analysis. After removing this subject, the correlation coefficients ranged from 0.38 to 0.92.

Analysis of variance: A three-way analysis of variance was performed over the judgments (repeated design). Results are given in Table 2. The experimental factors affecting the judgments are: S: subjects, R: resolution , M: metric and T: type of signals. All principal effects are significant (resolution: F(2,32)=195.0, p corrected < 0.01; metric: F(4,64)=16.3, p corrected < 0.01; type of signal: F(2,32)=41.1, p corrected < 0.00). Only the interactions between between *metric* and *type of signals* is significant at the lower threshold (F(8,128)=8.6 p corrected < 0.01). The principal effects of the experimental factors are depicted in Figure 5

[1]The stimuli used for the tests can be found at:
`http://www-sop.inria.fr/reves/projects/scalableAudio/`

[2]International Telecommunication Union

[3]*i.e.*, the subjects did not know which of A or B was the actual test or the reference.

Source	df	Sum of squares	Mean squares	F-value	p cor.
S	16	139.8	8.7		
R	2	967.7	483.2	195.0	0.000(**)
*S*R*	32	79.4	2.5		
M	4	23.4	5.8	16.3	0.001(**)
*S*M*	64	23.0	0.3		
*R*M*	8	5.6	0.7	1.9	0.191 (ns)
*S*R*M*	128	48.2	0.4		
T	2	112.7	56.3	41.1	0.000 (**)
*S*T*	32	43.8	1.4		
*R*T*	4	27.5	6.9	6.8	0.0191(*)
*S*R*T*	64	64.7	1.0		
*M*T*	8	24.5	3.0	8.6	0.010(**)
*S*M*T*	128	45.5	0.4		
*R*M *T*	16	17.2	1.07	2.8	0.115(ns)
*S*R*M*T*	256	99.1	0.4		
Total		1722.0	2.2		

df: degree of freedom
p cor.: corrected probability (conservative F-test)
* $p<0.05$; ** $p<0.01$; ns: not significant

Table 2: Anova table for the subjective evaluation

(vertical bars represent the standard deviation).

The bottom graph in this figure clearly shows the effect of the resolution on the average judgments: when 50% of the data are kept, average judgments lay between 0 and 1, almost meeting the requirement for transparency (i.e., no judgment above 1). At 25% resolution, average judgments rise to values between 1 and 2.5, and slide up to more than 3.5 for a resolution of 12.5%. The top graph in the figure indicates that musical signals were, on average, better ranked than the other type of signals (subjects freely mentioned during post-experimental interviews that differences were harder to notice for musical sounds). This indicates that the alterations of the signal produced by the algorithm are less perceptible for musical sounds. The middle graph in the figure represents the effect of the metric on the average judgments. Results were not quite as pronounced, but a first conclusion is that the A-weighting metric leads to the most audible difference between the processed and original sounds. Further understanding is obtained by studying the significant interaction between metric and type of signal, depicted in Figure 6.

The patterns of effects for the metrics are qualitatively identical for both musical and speech signals: A-weighting leads to the worst results, RMS level and Zwicker's loudness model result in the best judgments, Moore's loudness yields to slightly weaker judgments. On the other hand, for ambient sounds, Zwicker's loudness model results in the worst judgments, whereas Moore's loudness model leads the processed sounds to be better evaluated with reference to the original ones. Although, our evaluation was aimed at a totally different purpose, our results share some similarities with the recent paper by Skovenborg and Nielsen [26] that classified twelve loudness models (including several RMS-level metrics, Zwicker loudness and A-weighted level) into four categories. In their experiments, A-weighting was found to perform worst as a loudness metric while no clear advantage was found

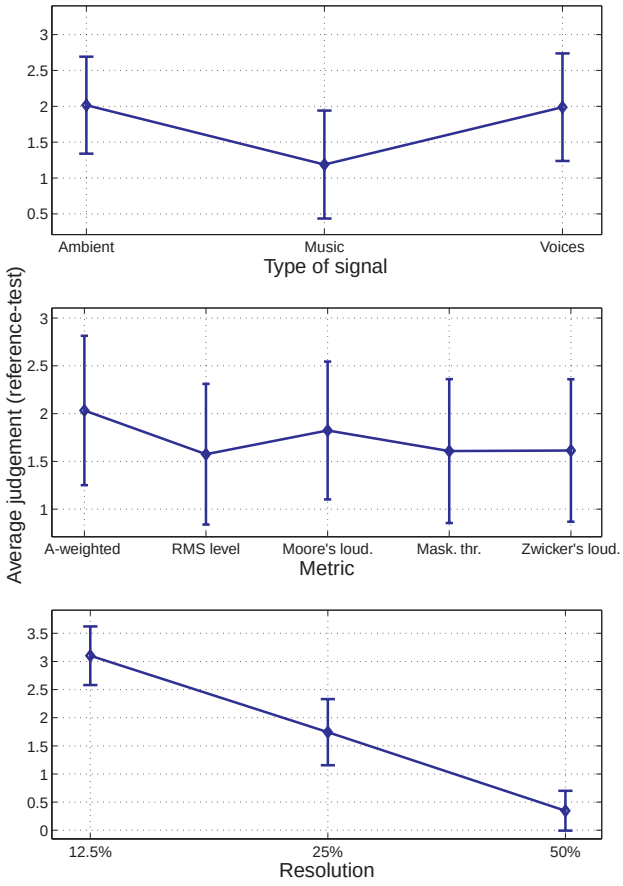

Figure 5: Principal effects of the experimental factors. Vertical bars represent standard deviation. The average judgment represents an annoyance level (hidden reference minus stimulus).

for the Zwicker loudness model over RMS-level related metrics. However, we could not test their two new loudness models, which seem to perform best. This would be an interesting future study to conduct.

5. DISCUSSION

Several conclusions can be drawn from this study. First of all, when only 50 % of the original data are used, subjects are almost unable to hear any difference between processed and original mixtures. When only 25 % of the sounds are preserved, average judgments lay between 2.5 and 1 (respectively "slightly annoying" and "perceptible but not annoying"). This indicates that our algorithm can reduce the required number of operations by more than 50 % without dramatically distorting the resulting mixtures (see Figure 7).
Another conclusion is that the judgments seem to be strongly influenced by the type of signal. However, as this variable also integrates several other effects (numbers of signals in the mix, sparseness of the mix, energy distribution in the mix, etc.) it would require further testing.

Nevertheless, differences between original and processed sounds were more difficult to detect for musical sounds. Two hypothesis

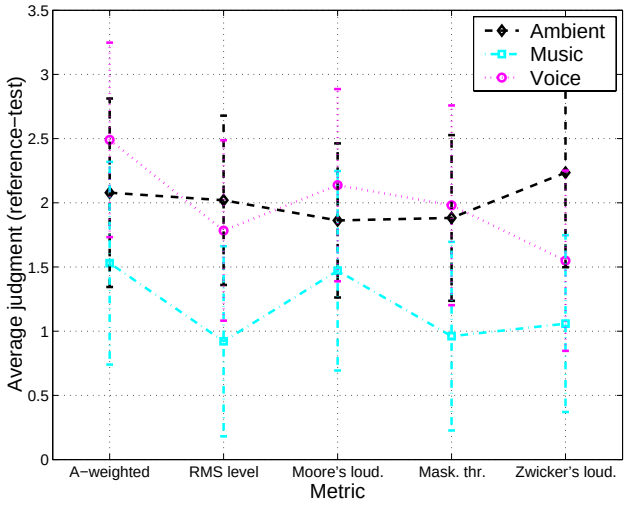

Figure 6: Interactions between the effects of the metric and the type of signal on the average judgments. The average judgment represents an annoyance level (hidden reference minus stimulus).

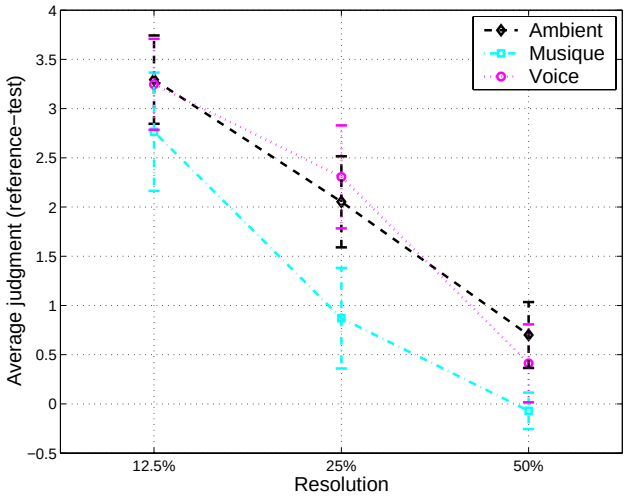

Figure 7: Averaged judgments for the three test mixtures and for three levels of detail. When only 50% of the input audio data was used, the resulting mixture was highly rated regardless of the stimuli.

can be formulated to explain this phenomenon: first of all, due to their nature, musical sounds are more sparse than other sounds. Energy peaks occur at regular rhythmic patterns, and there might a significant amount of low energy frames between these rhythmic accents. In our example, we estimated the sparseness (ratio between silence and signal) of our musical mix to be about 25%, which would make it well suited to our algorithm. However, the speech mixture was found to be much sparser that the ambient mixture (33% vs. less than 1%) although the results for these two cases were rather similar.

Another hypothesis is that the metrics were, in general, better suited to musical sounds.

Comparing loudness models, Zwicker's model leads to better results for speech and musical sounds, while Moore's loudness model performs best for ambient sounds. This is consistent with our implementation of Moore's loudness model for noisy signals (it can be reasonably assumed that ambient sounds are noisier than musical sounds).

These conclusions were confirmed during the interviews of the subjects. Many subjects reported that they used different criteria for the different types of signals. For speech signals, they reported to produce favorable judgments as long as the intelligibility was preserved, although most of the mixture was foreign language to them. For musical sounds, they did not hear any difference until the sounds were drastically distorted. Finally, for ambient sounds, they seem to have performed some kind of "spectral listening"; a typical remark being: "I tried to notice if there was more or less bass/treble". Hence, we can conclude that no metric seems to perform best in all cases but, rather, that the importance metric has to be adapted to the type of signal.

6. CONCLUSIONS

We have presented an approach for coarse-grain selective processing of audio signals. Several level-related metrics that can be used to drive the selection process were compared showing significant difference between the various metrics in terms of the ordering induced on the signals. A pilot subjective evaluation study suggests that A-weighting does not perform as well as the other metrics at prioritizing the sound signals. While RMS level appears as a good compromise, other metrics, loudness in particular, can yield to better results depending on the type of signals. Our selective processing approach integrates well within standard audio processing pipelines and can be used to reduce the necessary operations by 50% while remaining near-transparent or 75% with an acceptable degradation of the perceived quality.

As future work, we would like to explore extensions to finer-grain processing by combining our selection scheme with parametric audio coders or alternate representations for audio signals.

We believe that proposing and evaluating more sophisticated priority metrics is of primary interest for a wide range of applications including memory/resource management (e.g. 3D hardware voices, streaming from main storage space), real-time masking evaluation [6], on-the-fly multi-track mixing [27], dynamic coding and transmission of spatial audio content and more generally for computational auditory scene analysis.

7. ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their useful comments. This research was supported in part by the 2-year RNTL project OPERA, co-funded by the French Ministry of Research and Ministry of Industry.
http://www-sop.inria.fr/reves/OPERA.

8. REFERENCES

[1] B. L. Vercoe, W. G. Gardner, and E. D. Scheirer, "Structured audio: Creation, transmission, and rendering of parametric sound representations," in *Proceedings of IEEE*, may 1998, vol. 86, pp. 922–939.

[2] H. Purnhagen, "Advances in parametric audio coding," in

IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'99), New-Paltz, NY, 1999.

[3] J. Herre, "Audio coding - an all-round entertainment technology," in *Audio Engineering Society 22nd International Conference on Virtual, Synthetic and Entertainment Audio (AES'22), Espoo, Finland*, June 15-17 2002, pp. 139–148.

[4] H. Fouad, J.K. Hahn, and J.A. Ballas, "Perceptually based scheduling algorithms for real-time synthesis of complex sonic environments," *proceedings of the 1997 International Conference on Auditory Display (ICAD'97), Xerox Palo Alto Research Center, Palo Alto, USA*, 1997.

[5] W. Straßer and M. Wand, "Multi-resolution sound rendering," in *Symp. Point-Based Graphics*, 2004.

[6] N. Tsingos, E. Gallo, and G. Drettakis, "Perceptual audio rendering of complex virtual environments," *ACM Transactions on Graphics (Proceedings of SIGGRAPH'04)*, vol. 23, no. 3, 2004.

[7] M. Lagrange and S. Marchand, "Real-time additive synthesis of sound by taking advantage of psychoacoustics," in *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01), Limerick, Ireland, December 6-8*, 2001.

[8] K. van den Doel, D. K. Pai, T. Adam, L. Kortchmar, and K. Pichora-Fuller, "Measurements of perceptual quality of contact sound models," in *Proceedings of the International Conference on Auditory Display (ICAD 2002), Kyoto, Japan*, 2002, pp. 345–349.

[9] K. van den Doel, D. Knott, and D. K. Pai, "Interactive simulation of complex audio-visual scenes," *Presence: Teleoperators and Virtual Environments*, vol. 13, no. 1, 2004.

[10] A. B. Touimi, "A generic framework for filtering in subband-domain," in *Proc. of the Ninth DSP Workshop (DSP2000), Hunt, Texas*, 2000.

[11] E. Zwicker, "Dependence of post-masking on masker duration and its relation to temporal effects in loudness.," *Journal of the Acoustical Society of America*, vol. 75, no. 1, pp. 219–223, Jan 1984.

[12] F. Baumgarte, "A physiological ear model for auditory masking applicable to perceptual coding," in *Proc. of 103rd Convention of the Audio Engineering Society, , New York, USA*, 1997.

[13] R. Graham, "Use of auditory icons as emergency warnings: evaluation within a vehicle collision avoidance application," *Ergonomics*, vol. 42, no. 9, pp. 1233–1248, 1999.

[14] N. A. Stanton and J. Edworthy, "Auditory warnings and displays: an overview," in *Human Factors in Auditory Warnings*. Ashgate Publishing Ltd., 1999.

[15] N. A. Stanton and J. Edworthy, Eds., *Human Factors in Auditory Warnings*, Ashgate Publishing Ltd., 1999.

[16] Acoustics FAQ Section 8.1, "A-weighting formula," http://www.faqs.org/faqs/physics-faq/acoustics/.

[17] B. C. J. Moore, B. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness and partial loudness," *J. of the Audio Engineering Society*, vol. 45, no. 4, pp. 224–240, 1997, Software available at http://hearing.psychol.cam.ac.uk/Demos/demos.html.

[18] E. Zwicker, H. Fastl, U. Widmann, K. Kurakata, S. Kuwano, and S. Namba, "Program for calculating loudness according to din 45631 (iso 532b)," *Journal of the Acoustical Society of Japan*, vol. 12, pp. 39–42, 1991.

[19] T. Painter and A. Spanias, "A review of algorithms for perceptual coding of digital audio signals," in *Proceedings of the International Conference on Digital Signal Processing*, 1997, pp. 179–205.

[20] K. Brandenburg, "mp3 and AAC explained," *AES 17th International Conference on High-Quality Audio Coding*, Sept. 1999.

[21] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer, 1999, Second Upadated Edition.

[22] S. McAdams, M.-C. Botte, and C. Drake, "Auditory continuity and loudness computation," *Journal of the Acoustical Society of America*, vol. 103, no. 3, pp. 1580–1591, March 1998.

[23] D. C. Howell, *Statistical methods for psychology*, PWS-Kent, 1992.

[24] C. Grewin, "Methods for quality assessment of low bit-rate audio codecs," *proceedings of the 12th AES conference*, pp. 97–107, 1993.

[25] ITU-R, "Methods for subjective assessment of small impairments in audio systems including multichannel sound systems, ITU-R BS 1116," 1994.

[26] E. Skovenborg and S. Nielsen, "Evaluation of different loudness models with music and speech material," in *Proc. of 117th Convention of the Audio Engineering Society, San Francisco*, 2004.

[27] F. Pachet and O. Delerue, "On-the-fly multi track mixing," in *Proceedings of the 109th Audio Engineering Society Convention*, 2000.