DESCRIPTION-DRIVEN CONTEXT-SENSITIVE EFFECTS

Adam T. Lindsay & Alan P. Parkes

Computing Department
Lancaster University
[atl,app]@comp.lancs.ac.uk

Rosemary Fitzgerald

Music Department
Lancaster University
r.a.fitzgerald@lancaster.ac.uk

ABSTRACT

We introduce a new paradigm in digital audio effects that is based on more symbolic manipulations of elements of a sound, rather than using linear signal processing alone. By utilising content descriptions such as those enabled by MPEG-7, a system may apply context-sensitive effects that are more aware of the structure of the sound than current systems.

We advocate a standards-based approach (with MPEG-4, -7, and -21) so as to maximise the interoperability between different systems. The paper outlines MPEG-7 description structures that may be used as the basis for controlling and triggering effects in a system. It explores the different possibilities that are opened up by this paradigm. The way is then pointed towards more sophisticated control structures that may lead to more "musical" and dynamic effects.

1. INTRODUCTION

The vast majority of current digital effects are simply discrete versions of well-known analogue effects. Processing is essentially linear. In the cases where there is a non-linear model, e.g., granular synthesis, the processor uses little to no knowledge of the sound being processed. Modern computer-driven effects should take advantage of not only being *digital*, but also *computational*.

1.1. Multimedia standards

Recent and current developments in the MPEG standards are pointing the way to a new approach to manipulating digital content. MPEG-4 is hopefully well-known for the flexible signal processing available in the form of the structured audio orchestra language (SAOL). For the purposes of this paper, the general approach that MPEG-4 takes, that of an *object-oriented* audio codec toolbox, is more instructive.

A driving force in MPEG-4 development was the idea that much greater encoding efficiency could be gained by encoding individual audio and video elements with specialised codecs and then mixing the individual elements on the terminal [1]. For example, a vocal line could be encoded with a CELP (code-exited linear prediction) variant, a guitar solo with HILN (harmonic individual lines and noise), and backing drums and synthesisers synthesised with instruments built in SAOL. This aspect of the standard has been largely ignored in mainstream discussions of MPEG-4 in favour of the improved efficiency of component codecs (e.g., AAC). Still, the idea of sound being split into its component parts for further processing is a powerful and instructive one.

MPEG-21, the Multimedia Framework, is mostly recognised for its end-to-end digital rights management (DRM) strategy. However, more recent developments are revealing the overall plan for the framework to include more interesting capabilities than DRM alone. The Digital Item Adaptation (DIA) part of the standard [2] is in an immature state, but it shows some first steps at a standardised manipulation of pre-existing audio material. The stated goal of this manipulation is to adapt content for the terminal or the user, but reading through the current (April 2003) draft of the standard, it is hard not to imagine its rudimentary spectral and dynamic processing capabilities put to a more creative purpose. As of this writing, it is unclear where the standard will go in terms of exploring these possibilities, but it remains an interesting development, worth watching.

MPEG-7, the Multimedia Content Description Interface [3], offers the potential for a highly-detailed representation of many features in audio. The majority of its attention from the audio research community has been in its provision of a series of applications for audio analysis and machine listening systems. Most people view the description standard as enabling multimedia search and retrieval over the internet, but given a sufficiently-detailed description, it can be used for advanced audio-visual effects.

2. MPEG-7 DESCRIPTIONS

It is worth examining the various description structures available in MPEG-7 Audio. The structures that are of the most note to the audio effects world are segments, scalable series, time/frequency decompositions, and semantic labels of many kinds.

MPEG-7 segments are, most generally, partitions of an entire piece of media. Audio segments are contiguous temporal subsets of an audio media entity, demarcated with start and end times, generally based on one or more features. Segments are the fundamental unit for MPEG-7 audio descriptions: any instantiated audio descriptor has an audio segment as its ancestor in the XML-based description tree. They may have arbitrary temporal resolution. Audio segments can also be hierarchically decomposed, one segment containing any number of sub-segments. The described area within a segment may be discontiguous through the use of a "temporal mask" that excludes specific portions of the segment from descriptive consideration.

Within a segment at any level, there may be more detailed information, as embodied by the scalable series (see figure 1). The most common use of the scalable series is to offer regularly-sampled running values of a given descriptor, instead of an aggregate value for an entire segment.

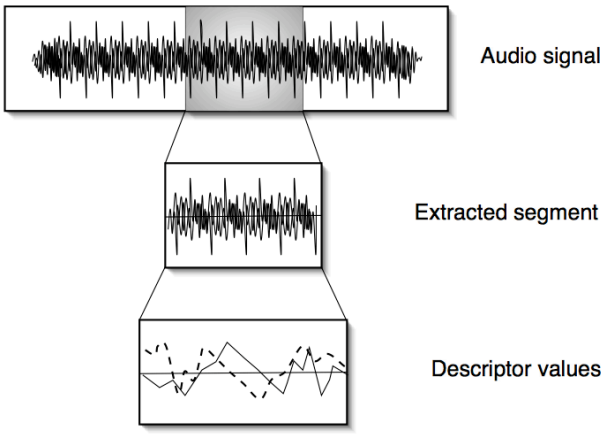

Figure 1: *A sound with multiple levels of description.*

Less a basic structure and more of a description representation, the sound effect description scheme is actually built around a general tool for audio analysis, including the spectrum basis descriptor. The general principle has been demonstrated for analysis and remixing of music [4], but it remains to be seen if the MPEG-7 descriptors, through resynthesis, offer enough fidelity to the original sound to be useful for audio effects.

The last MPEG-7 concept of importance to our discussion is the use of semantic labels of all kinds. Segments may have any number of labels attached, from phonemes and words to arbitrary thesaurus values. A phoneme-level segmentation and labelling could enable vowel matching and/or manipulation of a vocal line. Words from a controlled vocabulary might not only describe speech, but be used as labels for instruments or sound effects, enabling anything from triggers for events to replacement or proxy sound effects.

3. DESCRIPTIONS IN CONTROL

Rather than seeing content descriptions as being metadata for search and retrieval, we can view them as control elements. This requires examining fundamental decisions generally taken for granted in a more traditional effects model, broadly schematised in figure 2. A description may control *where*, *if*, and *how much* an audio effect is applied to a sound. The choice of *which* effect is probably left to an application or a user, but much of the parameterisation may be determined by a not-necessarily trivial mapping from content descriptors. Finally, it is worth examining the possibility that control comes from descriptions other than the sound that has effects being applied to it.

The first concern is *where* in the affected sound is being considered for effects. In this case, the description structures elaborated above are the most relevant factor. For a segment, the most fundamental locators are the start- and end-points. If available and appropriate, the time mask may come into play. If sampled descriptor values are available via the scalable series, then the precise time points may be used.

Given a candidate location in the sound where some form of effect may be applied, a system must determine *if* the effect should be applied. The most straightforward indicator is a label on a segment, allowing for a simple Boolean comparison. More sophisticated testing may be performed using numeric descriptor values, comparing them with a threshold or a range of values. An additional level of complexity can be gained by testing temporal behaviour or with values in another description.

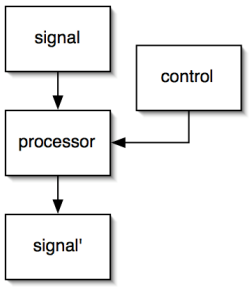

Figure 2: *A schematised version of traditional effects models. A signal is transformed into another signal' via a processing step.*

Once the affected portion of audio is selected, the *parameters* of a given effect must be chosen as well. The most basic of these parameters is how strongly the effect should be applied. Again, this can be directly mapped from descriptor values, or more complexly mediated by other conditions, such as by other descriptors, other descriptions, or non-linear, temporal functions of any of the above.

3.1. Description Structures

The *where* and *if* determinants above are fairly strongly dictated by the description structure used; segments and scalable series represent two distinct description philosophies that have respective advantages and disadvantages.

Segments are more in keeping with the object-oriented philosophy we are advocating. Each segment is an object, with its respective sub-description acting as parameters for manipulation. The object may have precisely delimited boundaries that are relevant to the manipulation in question. However, a segment must already exist in order to be manipulated. If the description does not contain the relevant segment-*cum*-sound-object, then it cannot be found. One only gets out what is put in. The semantics for describing the criteria for segmentation – and therefore the means for determining the relevance of a segment to any given search – is underdeveloped within the standard, and therefore must rely on best practice or other conventions.

Scalable series and sampled values in general are more flexible, but mean that descriptions are less pre-packaged as objects. The description generator does not need to pre-determine the segments of interest in a description, and the description consumer doesn't need to be prejudiced by the description's idea of relevant segments. Temporal resolution is determined by the sampling rate of the descriptors. The default within the standard is 100 Hz, but it can reach sample accuracy if desired. Search for sampled descriptor values based on arbitrary criteria can be very powerful, but can be more time-consuming as well, simply by dint of the number of comparisons that must be made.

3.2. Example

The canonical example of a simple description-driven effect is a compressor/limiter, driven by a continuously sampled AudioPower descriptor. One would first try to obtain a scalable series of AudioPower with a higher resolution than the default. With that in place, a very unsophisticated limiter could take those descriptor values and set them all to a single value, amplifying the underlying samples by a factor proportional to the factor needed to reach that single value.

Naturally, one can be more sophisticated in the *where*, *if*, and *how much* tests with the compressor. Not only can one set an arbitrary threshold and limit, but it is conceivable to set multiple "bends" in the compressor's curve. Setting attack and release times may be necessary for a more "natural" sounding compressor, but not necessary in terms of suppressing spurious deviations above the noise floor threshold. Since we assume that the description already exists, the processor may look ahead as far as necessary to be sure that the sound stays above the threshold.

With other descriptors added into the mix, one can add a degree of spectral dynamic processing. With a spectral centroid descriptor, this compressor could be made (for example) more responsive to high-frequency sounds, and ignore very low frequencies. Although a compressor/limiter is hardly a radical effect, hopefully the example has been instructive in how a subtle effect may be built from very basic description elements.

3.3. Multiplying possibilities

There are far too many possibilities to enumerate, but we give a range of aspects of descriptions that may result in different categories of effects. In our estimation, the most important factors affecting effects are structure, choice of descriptor, and source of control.

The structure has already been discussed in terms of the description structures available: the scalable series and the segment. Segments, however, may exist on many different scales, from the very large, architectural scale (e.g., cinematic act, movement), down to the building-block scale (e.g., phoneme, beat), and at any level in between (e.g., bar, phrase, section, syllable, word, sentence, subject/paragraph, etc.). The micro-level segments are particularly interesting because they provide an internal, synchronized clock in which the events themselves drive the effects. The macro-level segments are useful for architectural reorganizations and manipulations.

The choice of descriptors will affect the control of the effects in a fairly obvious way. They will most commonly be the parameters that affect the *if* of the control loop. We currently envisage the various spectral descriptors and labels (that draw from a controlled vocabulary) to be of the most use in the near-term, but specialized musical effects can be gained from the timbral and fundamental frequency descriptors as well.

The source of control is especially interesting. The most simple case is where the same descriptor that controls the effect is the parameter being modified by the effect, as in the simple compressor case described previously. A descriptor may, however, control any digital effect in one's arsenal: the fundamental frequency may be monitored so that when a singer reaches a high B, the flanger effect modifies the vocal track. The controlling descriptor need not describe the same sound that the effect is applied to; the flanger triggered by the high note might affect the drums and rhythm guitar instead of the vocals. Applying one description to another, such as the spectrum of one segment being modified to match another, would lead to a form of cross-synthesis. Any of these possibilities could combine with others, leading to a very wide palette from a relatively restricted descriptive vocabulary.

4. ISSUES IN DESCRIPTION-DRIVEN EFFECTS

4.1. Descriptions and Reflection

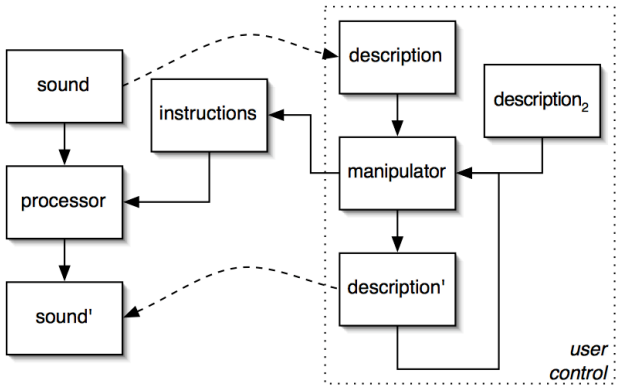

Figure 3: *A processing model in which the user modifies the descriptions, which is then reflected by the underlying sounds.*

The concept of control can be carried further when we consider descriptions to act as proxies for the sounds they describe. Given a sufficiently detailed description, an audio file itself is not needed until all of the effects are applied and/or the effects are previewed. This situation is not necessarily as irrelevant as the normal paradigm of an editor interactively and iteratively applying effects to a sound might suggest. Humans may be brought out of the loop, at least for more repetitious and tedious tasks.

By modifying the description so that its existing features are changed to desired values, then echoing those changes in the underlying sound, we approach the computational notion of *reflection* (see Figure 3).

4.2. Choices made in context

When manipulating the description itself to later be reflected in the audio, there are a number of issues to be tackled in terms of translating high-level control to signal processing algorithms. An example best illustrates some of the choices to be made by the user or the intelligent system.

A user selects a two-second segment from the middle of an audio file. She issues the command to make it 50% longer. How does the system prolong the segment by one second? It could do a very basic sample rate conversion, resulting in a slowed-down sound, with a pitch a perfect fifth lower. It could partially loop the sound, perhaps at a pre-determined point. There could be a spectral varispeed process, elongating the sound by the desired length, without changing the pitch. One could use PSOLA or similar time-domain techniques with

relatively simple sounds. Each of these techniques is plausible in one domain or another.

Another issue is how the expanded audio interacts with the surrounding file. In many cases, it is very conceivable that the entire audio file should be made one second longer. If this is not appropriate (for example, in the case of soundtrack editing), then the segment will overlap with its neighbours. It could be a centered overlap (overlapping both before and after the segment by 0.5 seconds), or before or after. With more sensitivity to the description, the overlap could be based on a hotspot; the sound is anchored at an extrema of a given descriptor (e.g., audio power) and is overlapped before and after that time point proportionally.

The issues raised by this simple example hopefully illustrate some of the complexities involved in granting a degree of context awareness to a description-driven audio effect.

4.3. Applications and use

We have concentrated on one paradigm, that in which the description and the sound already exist, probably generated by someone else, but fairly complete. It is an instructive model because it illustrates the fact that descriptions designed for one purpose, such as search and retrieval, can be used for audio effects as well. It also immediately suggests commercial applications, akin to the musical loop market now, in which audio with "pre-cooked" metadata is published for consumption and re-use by others. The metadata in the case of loops are simply loop markers, beats-per-minute, and perhaps key, but with descriptions, the metadata may extend to any audio descriptor, any label, or segment discussed above. A user may then use these heavily annotated source sounds in their effects engine of choice, and modify and assemble them at will.

Obviously, there are other possibilities. One is that the user may be unsatisfied with the quality, granularity, or flexibility of the source descriptions they obtain from a third party. This may lead them to generating their own descriptions, which in turn may lead to a more tightly-coupled description-manipulation loop. This in turn may lead to a continually evolving description as the processing engine calls for further analysis. The resultant descriptions would no doubt be very interesting in the variable level of detail, but would probably be of less generic use.

5. CONCLUSION

This paper has described a new paradigm for approaching context-sensitive, "intelligent" effects. The key driver behind the system's intelligence is the availability of detailed content-based descriptions of the audio to be affected. These descriptions feed into control switches and parameters for the effects.

Another major concern is with existing standards. The MPEG standards are of note because of their comprehensive approach to content description in MPEG-7, and because MPEG-4 and -21 are likely to be widely implemented. The widespread uptake of the predecessors of these standards also gives hope that the underlying infrastructure (in terms of descriptions available for sounds) is similarly widely available. We have limited our discussion to known descriptors within MPEG-7. Although it may turn out that the standards are currently insufficient for all desired features, they still form a good base upon which to build.

There is much work still to be done in the development of this model for processing audio, but hopefully we have set out some ideas that others can expand upon.

6. REFERENCES

[1] ISO/IEC, *Coding of Moving Pictures and Audio, Part 3: Audio v.2,* ISO/IEC 14496-3, Geneva, 2000.

[2] ISO/IEC, *Multimedia Framework, Part 7: Digital Item Adaptation,* ISO/IEC 21000-7, Committee Draft, 2003.

[3] ISO/IEC, Multimedia Content Description Interface, Part 4: Audio, ISO/IEC 15938-4, Geneva, 2002.

[4] Casey, Michael A. and Alex Westner, "Separation of Mixed Audio Sources by Independent Subspace Analysis," Proceedings of ICMC 2000, August 2000, Berlin.