AUDIOVISUAL CONCATENATIVE SYNTHESIS

Nick Collins
University of Sussex
Department of Informatics
N.Collins@sussex.ac.uk

ABSTRACT

Co-analysed audio and visual features can provide original concatenative synthesis effects in either or both modalities. This paper describes experiments in establishing audiovisual databases tagged by both audio and visual features, then creating new output streams by feature matching with a given input sequence; some intriguing effects are possible. Results are grounded in vectors of frame-wise data rather than higher level object recognition, but multi-modal processing extensions are discussed.

1. INTRODUCTION

Concatenative sound synthesis [7, 9] works with respect to large corpora of analysis-data-tagged sound segments, which can be recombined into novel outputs based [1] on target feature data streams. In most implementations, the database is pre-analysed, though it may also be formed on the fly; in either case, the target data can be from live input. Matches between the database and target are output after search. Questions of the efficiency of this search and the synthesis quality with respect to particular features and source databases are open research topics in current work.

Audio features can be leveraged for video analysis and cueing, and multimodal event databases formed on-the-fly during live performance [2]. The basis of the *sCrAmBlEd?HaCkZ!* project [2] is audio feature analysis on eighth note segments of a beat box audio signal, which are matched (by audio alone) against a database also containing video tags to pop videos. Video features can also be leveraged to extend audio analysis and cueing [6]. This is most familiar in current projects into automatic speech recognition, where an additional video component is applied for lip reading to disambiguate cases. [3]

This paper describes an exploratory study into the use of simultaneous audio and visual features for data-driven synthesis, from both technical and artistic perspectives. There is some overlap with engineering studies into multimodal data handling. For instance, Smaragdis and Casey [8] combine audio spectral amplitudes and image pixel intensity data into a single unified high-dimensional audiovisual feature vector, analysing this through Independent Component Analysis. In a recent study, Monaci et al. [4] construct multimodal dictionaries of learnt basis functions. Both these papers tend to assume that audiovisual correlation is present in the source material, which may not be the case for some artistic ventures. The features taken in this paper are perhaps more perceptually relevant than intensity value streams, though with a large reduction in dimensionality; yet they are still not at the level of semantic objects, and this is discussed at the close.

It should also be granted that a unified audiovisual view to features is implicit in the work of some current live audiovisual acts (klipp av, chdh, Coldcut's conception of audiovisual plug-ins for the FreeFrame plug-in format [4]), who may have artistic reasons to seek a greater equivalence or information transfer between modalities.

2. MODEL

Frame-wise and event-wise segmentation have been highly promoted in recent years (i.e. in music information retrieval) for operations on pure audio signals. The former is still the dominant trend, and whilst individual frames are typically at the level of FFT windows (10-20 msec), larger-scale processing suitable for auditory objects involves multiple frames at once; for instance, Casey and Slaney [1] promote the term *shingle* to describe the combination of M features per frame over the past N frames in larger M*N length feature vectors associated with a given frame.

In the case of this paper, each frame provides both audio and visual features. The 44100 Hz sampling rate used in this project provides an advantage related to its heritage in digital video - 30fps movie material approximates the audiovisual integration clock rate (30 msec period [5]), and corresponds to exactly 1470 samples. Blocks of 1470 samples (zero-padded to 2048 FFT blocks for spectral feature calculation) are taken for each video frame.

Whilst the joint feature vector for a frame is simply the concatenation of the audio and visual feature vectors [6], it is necessary to keep track of which feature is on which dimension so as to have independent weight control over their contribution to matching. It is also at the matching

[1] I might use the verb 'databased' in this instance!

[2] http://www.popmodernism.org/scrambledhackz/ This website has an excellent downloadable movie explaining the technical basis of the work.

[3] It should be noted that whilst such approaches can improve automatic speech recognition, they are not a necessary solution to the problem – for the blind are perfectly good language users, yet do not access any visual cues. Speech recognition is ultimately far more limited by contextual knowledge.

[4] http://freeframe.sourceforge.net/

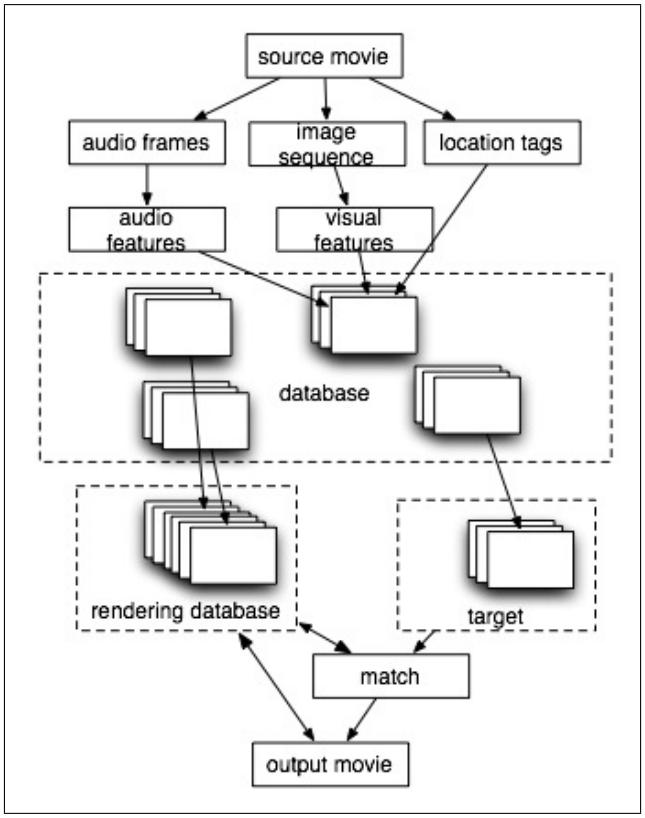

Figure 1. Batch processing view of audiovisual concatenative synthesis framework. The non-realtime case is exhibited here: the target (and also the rendering database via appropriate buffering) could also be the result of analysis on a live input stream.

stage that multiple frames are involved, by extending the distance calculation.

Image size is indirectly involved in the calculation of video features. For consistent output, the database just consists of movies at a common size (420x272 was a standard used). The target (input) sequence does not have to have the same size since the features are abstract descriptors of the motion or object content of image frames, independent of size through appropriate normalisation.

2.1. Audio and Visual Features

Five each of audio and video features were explored; more could easily be added, but a reasonably small number of features already empowers complex effects (for example, Bob Sturm's MATConcat uses six main (audio) features [9]). The added complication of audio and visual features gives plenty of scope for compositional effects to be discussed.

Most of the features are defined for a single frame, but change detection features were defined between frames for both audio and visual data, based on pointwise log differences. Tables 1 and 2 give the audio and visual features respectively – the features are (somewhat) analogous in pairs, though those of the images may deal with 2-dimensional spatial frequency rather than 1-dimensional

Feature	Implementation
specdropoff	log of the bin frequency for the 85% spectral power point
logpower	$\log\left(\frac{\sum x(i)^2}{1470} * (\exp(1)-1)+1\right)$
maxf0	log frequency for the peak (2048 point) FFT bin between bins 2-60
zcr	zero crossing rate: negative to positive crossings per frame
onset	change detector (see text)

Table 1. Table of Audio Features

Feature	Implementation
var	variance of the image
brightness	log of the average intensity of image pixels
maxf0	log frequency for the peak 2-dimensional FFT bin between bins 2-60 by 2-60
edge	edge detection count (see text)
imagediff	log difference pointwise between frames

Table 2. Table of Visual Features

spectra.

Before processing the image frames were converted to floating point descriptions in two matrix forms, the first retaining RGB channel separation in a three-dimensional matrix, and the second an average amplitude greyscale two dimensional image. The latter representation was used for the 2-D Fourier transform and edge detection.

Where the tables are sufficient for reconstruction I do not describe feature definitions in the text. For the audio, the onset detector is a variant of the family of bandwise log energy change detectors founded in the work of Anssi Klapuri (see [2] for further links). 96 subbands in the spectrum were taken, with linear spacing (every band) below 1000 Hz, and geometrically spaced bands over 1000Hz up to 11025; a single band encapsulated the top half of the spectrum. Subbands were formed by summing the amplitude spectrum for appropriate indices, then the log power was taken for that subband. Finally, only positive changes from the previous frame were summed as contributing to an attack percept.

A spatial zero crossing count was effected by edge detection [3, p. 175] using the Laplacian of Gaussian filter kernel, which finds zeroes of the second derivative of the smoothed signal, where the smoothing scale is determined by the parameter σ of a 2-dimensional Gaussian. The image filter coefficients are formed by the analytical solution:

$$\nabla^2 G_\sigma = \frac{x^2+y^2-2\sigma}{2\pi\sigma^4}\exp\left(\frac{-(x^2+y^2)}{2\sigma^2}\right) \quad (1)$$

For $\sigma = 1$ a 9 by 9 pixel kernel was calculated. 2-D convolution followed by a count of values close to zero (within +- 0.05) gave the edge detection count.

Figure 2 show the ten feature trails across 5 seconds of

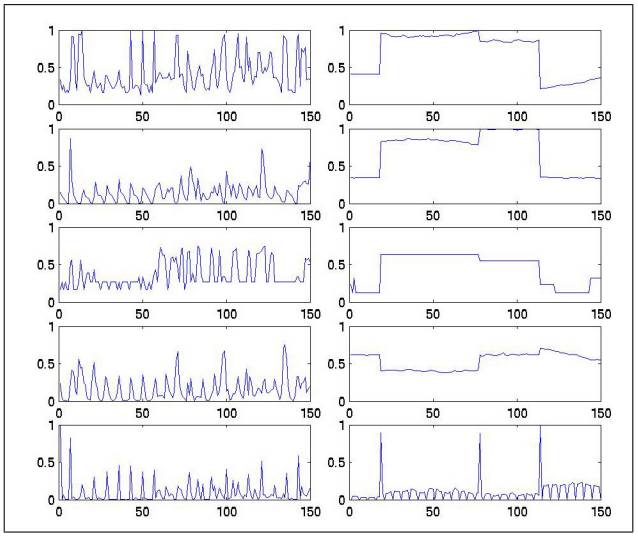

Figure 2. View of ten features across 150 frames (5 seconds) of a particular source; audio features on the left in the order top to bottom given by table 1, visual features on the right following table 2. Some correlations can be seen against the profiles – audio features appear noisier due to the greater time resolution of audio data, regardless of the framewise nature of calculation.

audiovisual source material. Some correlations between trails amongst audio and visual features (but not cross-modally) are seen. In the video the step functions are due to overt scene changes – these features are not discriminatory for semantic objects (as discussed at the close of this paper). Correlation coefficients for this situation show large correlations ($|r| > 0.5$) between the first and fourth ($r = 0.77$) and second and fifth ($r = 0.53$) audio features, amongst the first three visual features (all $r > 0.9$), and between the first and fourth ($r = -0.6896$) and third and fourth ($r = -0.7025$) visual features; but not between mixed audio and visual features. It should be made clear, however, that for other feature sets circumstances vary, and there is no evidence in general for a priori statistical dependence amongst the features. Such audiovisual correlation is discussed in greater detail below.

Ahead of feature matching, feature normalisation was carried out to obtain an equivalence of feature values (for otherwise, the ZCR count could be disproportionately influential compared to the log power, for instance). Normalisation by maxima and minima to the range [0,1] was the method followed in practice (the feature trails were empirically checked to make sure they did not produce significant outliers that would skew this mapping). Whilst statistical normalisation was investigated (correction to zero-mean and a std deviation of 1) the one-tailed or multi-peaked nature of the distributions on most parameters made this problematic.

2.2. Match procedure

Matching iterates through the frames of the target movie; each is matched to the best fit (with respect to conditions below) from the database. Following [7, p. 152], a weighted Euclidean distance metric is used for an extended target cost:

$$\text{cost of frame k} = \sum_{i=k-p}^{k} \sum_{j=\text{features}} (\alpha_j (x_j^i - y_j^i))^2 \quad (2)$$

Where x_j^i is the j^{th} feature of frame i in the database, y is the target feature set, and the α_j are the feature-wise weights. The 'shingling' is brought in at this point by variable p, the number of past frames to be considered in matching cost calculations.

Concatenation cost was zero; the default was thus an implicit trust in the extended multi-frame target cost as automatically finding appropriate matches in progression through the (usually reasonably continuous) feature vectors of the (self-consistent) target.

A brute force linear search was used for prototyping. Indeed, with the involvement of weights, it is hard to see what else could be done – new weights require recalculation [5] of a kD-tree or Ball-tree search or locality sensitive hashing map for approximate nearest neighbours [1]. The cost of feature calculation in the first place (particularly for images) is much greater than a linear search for best matches on the quantity of material considered.

3. IMPLEMENTATION

The prototype system was implemented in MATLAB, using Quicktime Pro to segment Quicktime movies into 30 fps JPEG sequences, and separate .wav audio files, suitable for easy loading and processing with MATLAB. Output from MATLAB were also JPEG image sequences and an audio file; these were re-combined again in Quicktime.

Because of the multiple stages in the process, and the cost of calculation (particularly for image features) the implementation was coded with independent rendering stages to promote quick experimentation. For instance, if a single new feature was added, this could be calculated for existing movies then imported at the synthesis stage, without affecting prior feature calculations. A subset of features and database material could always be used for rendering purposes.

4. AUDIOVISUAL EXPERIMENTS

Various demonstrations of audiovisual concatenative synthesis effects were constructed, and some brief notes on audiovisual effects obtained are now detailed. The source databases in the main consisted of downloaded movie trailers; cross-synthesis based on the latest releases breaches copyright if publicly released, but is fair use for research purposes, and definitely in the spirit of the alternative scratch video scene [2].

[5] To see this, imagine a weight tending to zero; this corresponds in the limit to projecting, such that widely divergent points on one dimension become packed closer together in the ensuing lower dimensional space.

4.1. Audio only

Standard audio concatenative synthesis effects are possible – but the original video frames associated with particular sound grains come along for free. This effect has been carried out with rhythmic pop video material with explicit eighth note segments in the *sCrAmBlEd?HaCkZ!* project. Rather than reproduce that, more abstract territory posited on shorter granular synthesis timescales (constructed from the 33 msec grains/frames) can be exploited.

4.2. Visual only

Here the audio is dragged along, and the matching proceeds only on the basis of visual feature analysis. The interesting byproduct is that the slower event rate of visual material (mediated in part by a slower scene/shot change rate) tends to promote the preservation of longer auditory segments; the output is less finely granulated, and this without the imposition of concatenation cost.

An abstract form of audio processing is possible by this mechanism, somewhat reminiscent of the use of complex numbers in equation solutions or hidden variables in physics: if the visual part is discarded, audio output can be generated indirectly from associated video content, the visual part of which is only utilised for processing, and never viewed!

4.3. Audiovisual/visualaudio

The choice of arbitrary weights amongst the features allows degrees of combination of audio and visual parameters in determining output. The more features are involved, the more difficult it is to predict the outcome. Indeed, some of the clearest output examples were generated with a low number of salient features.

An interesting twist is to vary the weights over time during rendering – this allows the matching to begin based on audio, interpolate through an audiovisual blend, and end up with entirely visual feature cued output.

4.4. Audiovisual correlations

The degree of existing correlation between audio and features depends on the nature of the source data. In the case of movie trailers used for the test data, correlation is intermediate – voiceovers and nondiagetic soundtrack music are uncorrelated (except in narrative content and emotional contingency) but sound effects are related for some audiovisual objects. Varying targets and database content vary the correlative factors. Just as for the features described in figure 2, it might be plausible to analyse correlation coefficients within and between audio and visual features ahead of rendering, to seek out a particular state of audiovisual correlation (where enabled by the situation) and set cost calculation weights appropriately.

5. AUDIOVISUAL OBJECT RECOGNITION

This paper has described prototyping of audiovisual concatenative synthesis from joint audiovisual feature vectors, at the level of the multimodal integration clock – 30 Hz, or blocks of 1470 samples (zero-padded to 2048 FFT blocks) at 44100Hz sampling rate.

One weakness of the current approach is the use of lower level features, rather than higher level structure; though a case can be made for the emergence of some higher level features implicitly from the time variation of low-level data. There are some critics that would not be content with anything other than semantic object recognition (though this can be less than well-defined). However, this preliminary study was an attempt to investigate novel compositional effects, and concatenative synthesis at a relatively low level remains an open area of exploration. Future work can enter a vaster world yet, of audiovisual object recognition, with higher level machine vision and listening processes for object identification running in both modalities. [6] Processes of data-driven material recombination founded on audiovisual analysis may have a very productive future as both technical tools and engines of novel aesthetic pursuit.

6. REFERENCES

[1] M. Casey and M. Slaney. Song intersection by approximate nearest neighbour search. In *Proc. Int. Symp. on Music Information Retrieval*, 2006.

[2] N. Collins and F. Olofsson. klipp av: Live algorithmic splicing and audiovisual event capture. *Computer Music Journal*, 30(2):8–20, 2006.

[3] D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, Upper Saddle River, NJ, 2003.

[4] G. Monaci, P. Jost, P. Vandergheynst, B. Mailhe, S. Lesage, and R. Gribonval. Learning Multi-Modal Dictionaries. *IEEE Transactions on Image Processing*, 2006.

[5] E. Pöppel and M. Wittman. Time in the mind. In R. A. Wilson and F. Keil, editors, *The MIT Encyclopedia of the Cognitive Sciences*, pages 841–3. MIT Press, Camb, MA, 1999.

[6] G. Potamianos, C. Neti, J. Luettin, and I. Matthews. Audiovisual automatic speech recognition: an overview. In G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, editors, *Issues in Visual and Audio-visual Speech Processing*. MIT Press, Cambridge, MA, 2002.

[7] D. Schwarz. *Data-driven Concatenative Sound Synthesis*. PhD thesis, Université Paris 6, 2004.

[8] P. Smaragdis and M. Casey. Audiovisual independent components. In *Proceedings of ICA*, pages 709–714, 2003.

[9] B. Sturm. Concatenative sound synthesis for sound design and electroacoustic composition. In *Proc. Digital Audio Effects Workshop (DAFx)*, 2004.

[6] Whilst the particular properties of audio and visual modalities should not be conceptualised as equivalent, there are some intriguing parallels (object processing/attentional resources to cope with three to four simultaneous objects, gestalt stream processing, logarithmic sensitivity) and differences (greater auditory temporal accuity, spatial properties, fundamental frequency, time of day based colour perception).