Audio-Visual Processing Tools for Auditory Scene Synthesis

Gavin Kearney¹, Rozenn Dahyot², and Frank Boland¹

¹Department of Electronic and Electrical Engineering, Trinity College Dublin, Ireland
²Department of Computer Science, Trinity College Dublin, Ireland

Correspondence should be addressed to Gavin Kearney (gpkearney@mee.tcd.ie)

ABSTRACT

We present an integrated set of audio-visual tracking and synthesis tools to aid matching of the audio to the video position in both horizontal and periphonic sound reinforcement systems. Compensation for screen size and loudspeaker layout for high definition formats is incorporated and the spatial localisation of the source is rendered using advanced spatialisation techniques. A subjective comparison of several original and enhanced film sequences using the Vector Base Amplitude Panning (VBAP) method is presented. The results show that the encoding of non-contradictory audio-visual spatial information, for presentation on different loudspeaker layouts significantly improves the naturalness of the listening/viewing experience.

1. INTRODUCTION

With the advent of High Definition (HD) formats offering enhanced visual imaging and superior audio quality to the cinematic experience, high resolution discrete sound reinforcement systems incorporating lossless coding technologies such as Dolby True-HD and DTS-HD have become commonplace, thanks to large capacity disc based mediums. Furthermore, sound reinforcement for motion pictures has spawned systems which now carry periphonic (with-height) information. The 22.2 surround system developed by NHK for example, utilizes three frontal tiers of loudspeakers for accurate reinforcement at the screen [1]. Smaller discrete formats such as the DTS-HD ‘Front High’ layout, utilize a standard ITU 5.1 setup with extra height information carried in elevated loudspeakers located just above the front left and right channels [2].

The aim of such sound reinforcement systems is to
enhance the video through good frontal imaging, rear effects loudspeakers and a low frequency effects channel. Although these systems were never intended for accurate soundfield reconstruction, developments in ‘G-Format’ Ambisonics decoding demonstrate how enhanced source imaging and soundfield reconstruction can be accomplished using irregular loudspeaker layouts [3, 4, 5].

Thus, as sound reinforcement moves from stable frontal imaging towards the ultimate goal of acoustic holophony, it is pertinent to re-address the question of accurately imaging the audio both horizontally and vertically to the video. In particular, we address the work implemented by foley mix engineers who incorporate everyday sound effects and background noises to create an auditory scene. In such cases if the panning automation data is drawn by hand, or even manually panned using a panpot, despite best efforts the audio-visual cues are likely to be inaccurate. Such processes are also tedious and time consuming.

To a lesser extent, we also address the necessity for surround mix engineers to have the vocal soundtrack relegated to the centre channel. It is noted that dialogue normalisation (dialnorm) is an important mastering consideration that requires mapping of the vocal to the centre. However, given the popular use of Automated Dialogue Replacement (ADR), where re-recording of dialog (‘looping’) is performed in a recording studio after the filming of the video, accurate vocal panning could lead to a better fit of the dialog to the on-screen actors, since the audio-visual cues would be non-contradictory.

The correct matching of the audio to the video in this regard is a valid concern since previous studies for stereophonic playback with television have shown that the auditory image should not vary more than ±6° to 11° [6, 7]. For cinematic rooms these limits become larger and can range between ±15° to 20° tolerance.

To this end, we present an integrated set of audio-visual rendering tools which will aid in acoustically segmenting and identifying sources, robustly tracking the visual positions of the sources, and synthesizing the audio according to this position and the chosen audio decoding scheme. We aim our approach to be applied not only to localisation of people in a video but usable for any object that has both a visual and auditory presence. For this, we believe that a user friendly tracker, easy to initialize and correct (when the tracker fails) and working in real-time, is a better option for supervised extraction of relevant trajectories, and ensuring a better post-processing tool for audio engineers.

We will begin by presenting an overview of our integrated processing tool. We will then introduce the video tracking, audio segmentation and enhanced audio spatialisation schemes utilized. Finally, we present the results of a subjective test of several original vocal soundtrack mixes against auditory scenes created with our synthesis tools.

2. SYSTEM OVERVIEW

Audio post-production typically comes after the majority of the visual post-production and editing is done, and usually at a small fraction of the overall film budget. The edited film is segmented into shots (continuous sequences of frames) and the sound engineer has several separated audio stems (sub mixes of dialogue, sound effects, music, etc.) to typically mix into stereo, 5.1, or 7.1 audio formats to match the video. In order to encode the spatial location of the audio sources, the mix engineer must first identify the visual object with which to acoustically position the audio source. To facilitate this audio-visual object segmentation and panning, we have developed several tools. These are summarized in two different processing chains in Figure 1.

First, the visual object of interest is selected in a frame of the sequence (e.g. in the first frame of the shot). This task can be done by selecting the region of interest by hand using the mouse. When considering standard visual objects such as faces, this task can also be performed with an automatic detector.

As a parallel process, segmentation on the audio stream can be performed automatically to separate the events of interest. This task can be done by selecting the region of interest by hand using the mouse. When considering standard visual objects such as faces, this task can also be performed with an automatic detector. The second process on the visual stream implements tracking of the object until the end of the shot if necessary. If the object is lost, because of occlusion for example, the engineer has the possibility to intervene to re-initiate the tracking where and when it was lost.

As a parallel process, segmentation on the audio stream can be performed automatically to separate the events of interest. This is useful when, for example, a single vocal stem is used for all the dialogue. The grouping of the audio events with the
Fig. 1: Processing tools for audio-visual object segmentation (the computer icon signifies an automatic task whereas the screen bean indicates human interaction).

The corresponding visual object is done by hand by the sound engineer.

Once the object of interest has been grouped with its chosen audio, two possibilities exist for audio resynthesis for spatialisation (B-Chain in Figure 1). The first is the generation of automated panning metadata to be used with a VST host. The second is the actual render of the audio to a given spatialisation scheme. Both are dependent on the playback parameters given by the engineer, which specify the screen size used and the loudspeaker layout.

Having given an overview of the basic operations of the software, we can now explore in more detail the methods employed in both the A and B Chains.

3. A-CHAIN PROCESSING

The A-chain of the software is concerned with determining where and when the visual object to be tracked is, as well as its corresponding audio track. Characters and their dialogue are of principle interest here, where faces (visual objects) have to be associated with their corresponding voices (audio objects). Both these objects need to be localised in their respective data streams.

3.1. Object tracking

Dedicated techniques such as face detection, face recognition and tracking, can be used to index videos according to the characters appearing in it. This can be possibly achieved with the additional help of textual transcript and subtitles [8] or audio information [9]. Good performance, however, depends on an exhaustive learning process to allow recognition despite changes in the object appearance.

The implemented software uses several functionalities available in the Intel OpenCV library [10] developed for real-time applications in video processing. In particular, the video is displayed on screen and several user friendly controls allow the activation of the following functionalities:

- **P**: Play/pause, the first frame is display in pause mode.
- **R**: Rewind to go backward.
- **F**: Automatic face detection [11]. Since faces are major objects of interest, face detection has also been made available in our software. This detector is based on prior training of positive and negative examples of faces using Adaptive Boosting (AdaBoost) on Haar like features [11].
- **M**: Automatic mouth detection [11]. The face detector can be trained to detect other classes of objects. We therefore added a mouth detector.
- **Right mouse button click**: Point selection. Here the selected corners are tracked [12], and the mean and covariance of the selected points are used as an estimate of the object position and size. An example of tracking using this technique is shown in a recorded sequence of a lecture in Figure 2.
- **Left mouse button held**: Region selection. The region is tracked through the colour histogram, with the CAMSHIFT (Continuously Adaptive Mean Shift Algorithm) algorithm [13]. CAMSHIFT tracks the center and size of the probability distribution of the colours of an object of interest selected on a still frame at a given time in the video.
It should be noted that the performance of the detectors is dependent on the position of the actors as well as the quality of the video. Automatic detectors can help in many cases, but selection by hand may be as equally efficient for selecting objects such as faces. Moreover, the software can track not only people, but also any other object generating noise (animals, cars, explosions, etc.). Both point detection and CAMSHIFT can be used simultaneously, increasing the robustness of the estimation of the object trajectory. The A-Chain gives the spatial-temporal track of the visual object as an output.

![Fig. 2: Video tracking using feature points](image)

(a) Initialization (b) tracking (c) Outlier point (d) User correction

3.2. Audio segmentation

Once a visual object has been successfully tracked, it is necessary to identify and segment the acoustic source of the object for the soundtrack. If the source is located in a mix ‘stem’ file then segmentation can be achieved either automatically or manually. Automatic segmentation of the audio is performed by calculating the envelope of the waveform and allowing the user to specify a segmentation threshold, as can be seen in Figure 3. A compensated attack and decay time can also be implemented. Source detection in the soundtrack segment can then be derived by comparison of a cepstrum representation of the audio to predefined training sequences [14].

![Fig. 3: Automatic segmentation of audio stems](image)

3.3. Matching of Audio and Visual objects

 Unless thorough priors are known by the system, there cannot be automatic association in between one particular sound with one specific visual object. In the case of dialogue, a speech fragment can usually be associated to one particular visible actor if his/her lips are moving in the image sequences synchronised with the sound. Cutler et al. [15] have proposed to use correlation in the audio-visual streams to perform this audio visual matching automatically in video conferencing. However, their approach for lip detection is only suited for videos without camera motion. Detection of lip movement in films is generally not an easy task when considering all possible difficulties encountered in edited movies such as lip occlusions, low resolution of lips, camera motion etc. Additional problems occur when the audio object, sound or speech, do not match any visual object on screen. Currently, this correspondence is implemented by hand.

4. B-CHAIN PROCESSING

Where the A-Chain is associated with audio-visual tracking, the B-Chain is concerned solely with the spatialisation of the audio. The inputs to the B-Chain are the segmented audio sections with their corresponding spatial-temporal metadata. The engineer must then decide which rendering scheme is to
be used to spatialise the audio as well as specify the loudspeaker layout and corresponding screen size for rendering. If rendering is not desirable the spatial-temporal metadata is output as intensity panning automation data, reformulated for the given dimensions.

4.1. Rendering Scheme

For a given loudspeaker layout, it is necessary to choose an appropriate spatialisation scheme with which to render the audio. Recent studies by the authors assessed the subjective and objective performance of various sound spatialisation techniques using an 8 channel array [16]. The techniques were assessed in terms of their localisation accuracy for virtual sources, for a distributed audience, in a reverberant environment. The systems tested were Delta Stereophony System (DSS), Vector Base Amplitude Panning (VBAP) and Ambisonics. The results showed that source localisation for non-central listener positions is consistently biased away from the intended image position, irrespective of the spatialisation technique tested, or the nature of the source stimulus. Vector Base Amplitude panning (VBAP), however, is not as affected by off-centre image shifts to the same degree as the other spatialisation systems, due to the smaller number of contributing loudspeakers (two or three), and gives the closest average match to the intended position of the virtual sources [17].

One possible solution to these off-centre image shift problems can be found in systems which attempt to synthesise the accurate wavefronts required for localisation for a distributed audience. The Wave Field Synthesis (WFS) approach, proposed by Berkhout et al [18] is one such solution. However, this system is currently impractical and expensive for both cinematic and domestic implementation.

Currently the implemented software provides support for VBAP and Delta Stereophony implementations of 2.0 and 3.0 stereo, ITU 5.1 and 7.1 ‘Front-High’. It should be noted that the general audio-visual rendering principles outlined in this paper can be applied to any spatialisation scheme.

4.2. Screen Dimensions

An excellent overview on the geometrical assignments of image width to listener position and viewing angle for high definition formats is given in [19]. For appropriate audio-visual matching it is highly important that the corresponding listening and viewing angles are non-contradictory. For example, the recommended viewing angle for HDTV (1920 x 1080p) is 30° at distances of 2H-4H (H = times the screen height). For film with very wide angles (pixel resolution of 3840 x 2160) it’s 45° at a viewing distance of 1H-4H. The contradiction of the recommended listening angles at ±30° has been implemented in the software, and the engineer can specify both the aspect ratio, the listening distance and the listening angle. The stereophonic image is then adjusted to accommodate the desired format for the centre listening position.

5. SUBJECTIVE ASSESSMENT

Subjective tests were created to compare different auditory renderings of several video segments. The tests were conducted in a listening room in Trinity College using the speaker layout shown in Figure 4. The speaker configuration was chosen so as to allow direct comparison between soundtracks played over stereophonic, standard ITU 5.1 and DTS-HD (with height) 7.1 layouts. In accordance with the ITU-R BS.1284-1 recommendation for listening tests, 10 expert listeners, were chosen for the tests [20]. Each listener was under 35 years of age, of excellent hearing, and highly experienced in musical production. The loudspeakers were calibrated to 79dBA at the listener position. The aspect ratio for each of the clips was chosen as 16:9 and the basis width of the image adjusted for a vertical viewing angle of ±24°. The bottom LCR loudspeakers were adjusted to ear height (1.1m) and the height loudspeakers were raised to accommodate a 27° elevation view. The stereophonic imaging was correspondingly adjusted for these angles and an acoustically transparent screen was used for the high definition projection. The spatial rendering method employed was VBAP.

In the tests, each participant was individually tested and was presented with three different video clips: one from a commercially released movie, one from an animated sequence and one from footage of a lecture. All three clips contained dialogue. For each video clip, the listeners had to compare our audio enhanced sequences to the original ones (presented in random order), in three different audio formats.
(stereo, 5.1 and 7.1). The participant was allowed to request repeat listening of any of the two segments that were presented to them for a given format. All presentations were 96kHz, 24-Bit.

5.1. Tested videos

The first video sequence used was a shot from a commercially released movie. This shot corresponds to a dialogue in between two characters on the left and right of the screen. There are attempts to provide a reasonable spatial impression by the original mix engineers, but the dialog suffers from obvious studio ADR. The second sequence is an animated scene, again corresponding to a dialogue in between two characters, with one of them moving from right to left while speaking. The final sequence is an amateur video simulating a lecture environment. Some images are shown in Figure 2. Here a lapel mic was used to record the audio, making a significant challenge in constructing the auditory scene for ‘natural’ presentation.

5.2. Questionnaire & Data Analysis

In a questionnaire running concurrently with the tests, each participant was asked to rate the naturalness of the audio in relation to the video on a scale of 1 to 5 where 1 is bad and 5 is excellent. This was used to gauge the perceptual effect on naturalness of panning with ADR and unnatural sounding audio. Secondly, the participants were asked to judge the coherency between the audio and video for each presentation. This allowed us to examine the transparency of the audio-visual tracking. Finally the overall quality of the audio was examined, in order to investigate any perceived differences due to cross-modal interaction and non-audio-visual contradictions.

6. EXPERIMENTAL RESULTS

In the following, original refers to the original recording in stereo, 5.1 and 7.1; enhanced refers to our enhanced mixes, where spatial localisation is encoded in the sound. The answers given by the listeners have been analysed in the following paragraphs to assess which mix was preferred to the other. In fact, only three ratings were used: A better than B, B better than A, and both the same.

6.1. Naturalness

The question asked in the questionnaire was: “Which of the two movie clips has the most natural sound?” Three videos were used, and three mixes were compared: stereo vs enhanced stereo, 5.1 vs. enhanced 5.1, and 7.1 vs enhanced 7.1. The results shown in Table 1 show that the enhanced version of the mixes is found to be significantly more natural. Low percentage in favor of the enhanced mix indicates high indecision in the answers. For instance, 50% of the listeners preferred the enhanced 7.1 mix in the sequence lecture, 20% were undecided (they found that both sequences were the same), and the remaining 30% preferred the original 7.1. In Fig-

<table>
<thead>
<tr>
<th></th>
<th>Stereo 70% 73.3%</th>
<th>5.1 70% 70%</th>
<th>7.1 70% 6.7%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movie</td>
<td>80%</td>
<td>70%</td>
<td>70%</td>
</tr>
<tr>
<td>Animation</td>
<td>50% 10%</td>
<td>70%</td>
<td>90%</td>
</tr>
<tr>
<td>Lecture</td>
<td>66.7% 6.7%</td>
<td>80%</td>
<td>70%</td>
</tr>
</tbody>
</table>

Table 1: Percentage of people who found the sound more natural in the enhanced mixes (main figure), and percentage of people who found both clips equivalent (exponent).
all the sequences and all the formats. We see that the enhanced version of the mixes have been chosen at 70% vs. 22% for the originals, and that all the results are similar whatever the sequence played and whatever the format. These results indicate that despite conventional panning methods, in particular for dialogue, a more natural presentation can be achieved by having non-contradictory audio-visual cues.

6.2. Spatial Coherency of audio-visual streams

Listeners were also asked which of two sequences has the most coherent audio and visual streams for localisation. In other words, how well did they feel the audio matched the on-screen visual? Table 2 shows that our enhanced mixes have been chosen above the standard mixes. High correlating values were found for the stereo enhanced versions, indicating that such auditory enhancement could be highly applicable to HDTV formats with stereo reproduction. In particular, the amateur lecture video sequence, with its unnatural sounding lapel microphone audio achieves 90% of the votes. If this result is coupled with the 80% vote for naturalness for this clip (Table 1), then it can be seen that spatial coherency can help overcome such unnatural sounding audio. This could be particularly useful in teleconferencing for example.

Figure 6 shows the overall average spatial coherency in between the audio and the video streams as perceived by the listeners. Over all the three types of sequences tested, and considering all the audio mixes (stereo, 5.1 and 7.1), an average of 72% of the listeners perceived the enhanced mixes as better for spatial localisation. On average 10% were undecided (i.e. on average only 20% of the listeners preferred the original mixes). This again strongly correlates with the results of Figure 5, indicating that correct audio-visual matching and a sense of naturalness are closely linked.

6.3. Audio quality

Finally, listeners were asked to rate the audio quality of each sequence. Quality here, refers to the actual fidelity of the audio as well as the comfort of listening. As Table 3 shows, no significant differences were found, with listeners rating the overall quality unchanged between each sequence pair. This, in itself is a significant result, since it shows that no discomfort was found with our enhanced sequences, despite the fact that listeners are already comfortable with conventional panning methods. Thus, an increase in naturalness using our method does not translate to a decrease in the comfort and fidelity of the overall audio.

Figure 7 shows the overall average audio quality perceived by the listeners over all audio-visual clips.
Table 3: Percentage of people who found the audio quality better in the enhanced mixes and percentage of people who found both clips equivalent (exponent).

<table>
<thead>
<tr>
<th></th>
<th>Stereo 5.1</th>
<th>7.1 66.7%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movie</td>
<td>20%53.3%</td>
<td>26.7%46.7%</td>
</tr>
<tr>
<td>Animation</td>
<td>20%66%</td>
<td>30%60%</td>
</tr>
<tr>
<td>Lecture</td>
<td>20%46%</td>
<td>30%40%</td>
</tr>
</tbody>
</table>

Again, no loss of quality is felt in between the original and the enhanced mixes, and both are assessed to be the same at 54%, with the remaining votes equally shared between the original and enhanced mixes.

Fig. 7: Overall perceived audio quality: enhanced (blue), original (pink) and same (yellow).

7. CONCLUSION

We have presented here an integrated set of audio-visual rendering tools to aid the audio engineer in soundtrack mixing for film. It was shown through subjective testing that the selected methods used for video tracking and auditory segmentation and synthesis are effective in creating enhanced mixes. Furthermore, with non-contradictory audio and visual localisation cues it is possible to significantly enhance the naturalness of audio-visual presentations without affecting the overall subjective quality. Further work will focus on a VST/RTAS implementation of this software for use in post-production sequencer packages. An analysis of the rendering methods will also be implemented for audio-visual musical presentations as well as off-centre listening positions.

8. ACKNOWLEDGEMENTS

The authors would also like to thank the volunteers for the subjective listening tests, as well as DTS consultant Jeff Levison for his advice on the DTS-HD standard. This research is supported by Science Foundation Ireland (SFI) and Enterprise Ireland (IP-2006-0412).

9. REFERENCES


