Automatic sonification of video sequences through object detection and physical modelling

Master Thesis

Andrea Corcuera Marruffo

AALBORG UNIVERSITY
DENMARK

Aalborg University Copenhagen
MSc. Sound and music computing
2017

Abstract

The sounds generated by the objects that surround us is intrinsic to our life. We associate some specific sounds to the objects, its characteristics, and the actions that generate them, and we expect to hear the corresponding sound when we see that item. Similarly, one expect to see the corresponding object when we hear its sound. For example, on the street, when we hear a characteristic sound of a motor, we know that a car is approaching. There is, therefore, a particular relationship between the objects and the sound that they produce.

In films, many of these sound effects are added in post-production, a method called "Foley". In this project, these sound effects will be generated automatically based mainly on one characteristic of the objects involved: their material. A system based on an object detector, an impact detector and a sound modeler will be presented. A perceptual evaluation in which the subjects will watch some videos and listen to the sounds predicted by the model will be performed. In this way, it will be tested if the sounds proposed by the model are played in accordance of what they expect to hear.

Contents

Chapter 1

Introduction

1.1 Sound in films

From the beginning of cinema, numerous attempts have been made to merge sound and visual content in order to offer a better experience to the audience. Initially large orchestras were employed in the theaters to play music along with the movie. In smaller venues, where they couldn't afford big orchestras, a pianist accompanied the silent film. Adding sound to the movies was, therefore, a big expense since they had to hire musicians to play every time the movie was displayed. Thus, the focus was put on getting pre recorded music to play with the film.

The first film which that included dialogues and music was "The Jazz Singer", released in 1927. This musical film was produced with the Vitaphone system [1] and is commonly considered he first talkie film. It was very primitive, it consisted of a disc and a projector, so timing errors occurred often. Synchronization between sound and film was later guaranteed by recording the sound in the same strip of film that contains the pictures. This became the standard until the advent of digital revolution.

The audio of the video is usually recorded at the time the video is filmed but other *diegetic sounds* (those whose source is visible or is implied to be present in the film) are often added afterwards. While filming, most likely it won't be possible to capture all the sounds involved in the scene, won't have the desired quality or they won't be very audible. In movies, the boom operator records the dialogue and avoids to record other sounds, some of which are added in post production. By adding these sounds effects after filming, the audio engineer has total control over their quality and relative intensities.

These sound effects that are added to videos and movies in post-productions are called Foley sounds. This technique, introduced by Jack Foley in the 1920s and that uses a large variety of objects to imitate or create sounds, has been widely used up until now. These kind of sounds are usually added to the original audio recordings but in other cases, all the sounds of the film have to be generated afterwards, like in animated movies. They add a deeper sense of realism and improve the auditory

experience of the movie.

Many of these sounds are *impact sounds*, produced by the collision of objects and characterized by their short duration, abrupt onset and quick decay. These sounds give us information about the properties of the objects involved as well as the action that has generated them. In real life, we can fairly predict the sounds that will be produced when some actions are performed, and we expect to hear their corresponding sounds when we see these actions. It is important, therefore, that in audiovisual contents video and audio are strongly linked to reality unless another perspective is desired, like, for example, in cartoons.

1.2 Motivation and goals

Foley sounds are usually performed by Foley artists in a Foley studio, with a wide variety of objects and surfaces. They create the appropriate audio effects while watching the footage they are going to add sound to.

A cheaper way to get these sounds is to use some pre-recorded audio effects, without the need of paying any Foley artist to create them. At the present, there are many libraries that offer a large variety of sounds ready to use in the videos. The sound designer has the task of choosing a specific audio file and synchronizing the action in the movie with the pre-recorded sound.

The timing has to be perfect; the sounds, which complement or substitute the original ones recorded at the time of filming the video, must be synchronized with the action that the audience is looking at. It can be a very time-consuming work, and therefore some simple actions could be sonified automatically to make the job of the sound designer easier and faster.

In this thesis a solution to this situation is presented. The algorithm proposed synthesizes automatically the audio generated by a constant source of sound (for example an engine) as well as the impact sound effects of rigid bodies. In addition, it also locates the sound source and makes a stereo panning of it.

However, unlike traditional work on this problem [2; 3; 4] that focus on 3D models, in this thesis the parameters needed to generate the sound are extracted afterwards from the 2D video recording. This is possible thanks to the use of a convolutional neural network which is able to detect the objects present in the video image. It labels the objects of the scene and provides information about their relative position in the image. Using these parameters, the sound is automatically synthesized afterwards using physical modelling synthesis.

1.3 Structure

The next chapters of this thesis are organized as follows: in chapter 2 a review of different techniques for sound synthesis can be found and in chapter 3 some traditional methods for object detection are presented, with an emphasis in convolutional neural networks. The description of the application developed is included in the chapter

4, the chapter 5 contains the description of the conducted evaluation together with the discussion of the results, and finally the conclusion of this thesis is contained in chapter 6.

Chapter 2

Sound synthesis

2.1 Introduction

Sound synthesis is the process of generating sound without using any acoustic instrument. This synthesis can simulate musical instruments, natural sounds, or even create new sounds.

Some concepts used in sound synthesis are explained as follows [5]. A *digital oscillator* is a sound source that repeats a waveform with a specific amplitude and fundamental frequency. The most common waveforms used are the sine wave, the sawtooth, the triangle and the square waves. The source of the sound samples in the oscillator may come from a mathematical formula or from a *wavetable* that is built beforehand. Instead of generating sound samples by computing mathematical operations, the source in this case is an array with N equally spaced points that depicts one whole cycle of the oscillator. Each value in the table represents a signal amplitude at a particular point in the cycle. The signals that are generated by the sound source can be modified by a *filter*, that alters the magnitude of the frequencies of the input by letting pass some frequencies and attenuating others. The filter may change over time and controlled with an *envelope*. Some popular techniques that make use of these modules are shown below.

2.2 Sound synthesis techniques

2.2.1 Additive synthesis

The additive synthesis, one of the first computer-music synthesis methods, generates sound by adding different sine waves [6] .

It is based on the Fourier theory that states that any periodic signal is made up of a sum of multiple sinusoids. Each sinusoidal component may have a different amplitude and phase envelopes that change over time.

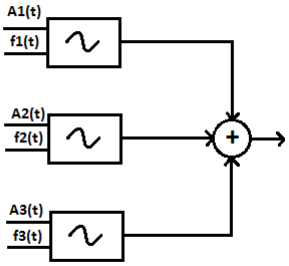


Figure 2.1: Additive synthesis

2.2.2 Subtractive synthesis

This method does the opposite of the previous one. It attenuates unwanted elements of a complex signal (usually with many frequencies) to generate the desired sound. It uses digital filters to alter specific frequency components of the sound and adjust it to the spectrum of the desired result. The sound sources used to excited the system can range from impulses, periodic train of impulses to noise, and the filters used can be from simple narrow filters to more sophisticated filters with time-variant coefficients. [7].

2.2.3 Modulation methods

Modulation synthesis is the modification of the amplitude, frequency or phase of a simple periodic signal (the carrier) by another signal (the modulator).

Amplitude modulation

In this technique, the amplitude of the carrier is modulated by another signal. It is mathematically expressed as:

$$y(t) = cos(2\pi f_1 t)(cos(2\pi f_2 t) + K) \tag{2.1}$$

Where f_1 is the frequency of the sinusoidal carrier and f_2 is the modulator frequency. If $K = 0$, it is equal to **ring modulation**, which differs from amplitude modulation in the components of the result, since in ring modulation the frequency of the the carrier signal doesn't appear in the spectrum.

Frequency modulation

This technique is based on the modulation of the frequency or phase of a simple periodic waveform with frequency f_c (the carrier) with another simple periodic waveform with frequency f_m (the modulator). This causes diverse sideband sinusoids with frequencies derived from the carrier frequency plus and minus integer multiples of the

modulator frequency [7; 8]. It can be expressed as:

$$y(t) = sin(2\pi f_c t + I sin(2\pi f_m t)) \tag{2.2}$$

Where I is called the modulation index and determines the strengths of the Kth side components given by the Bessel functions of Kth order. A rule of thumb, Carson's rule, considers that the number of significant sidebands is approximately equal to $I + 2$.

2.2.4 Physical modelling methods

Physical modelling techniques aim to generate a waveform by using mathematical approximations that simulate the physical processes that produce the sound in a real acoustic instrument or sound event.

Modal synthesis

Modal synthesis states that the sound produced by a vibrating object can be generated by the sum of their modal components (particular patter of vibration that is associated with resonances in the spectrum).

The acoustic response of an object to an impulse at location k can be seen as a sum of modes (damped sinusoidal waves):

$$y(t) = \sum_{n=1}^{N} (a_{nk}) e^{-d_n t} * \sin(2\pi f_n t) \tag{2.3}$$

where $a_n k$ is the amplitude of the mode at the location k and f_n is the frequency of the mode n. The damping coefficient of each mode is represented by d_n and is highly influenced by the material.

This technique is an efficient way of generating the sounds of objects that show a relative small number of main modes.

Modal Synthesis can be used to generate complex interactions as well. Van den Doel et al. introduced an algorithm that used modal models to automatically simulate sounds of impact, sliding and rolling [2] for interactive simulations as games and for animations. In [9], Ren et .al present a method that uses pre-recorded impact audio clips to estimate the material parameters associated to some particular objects. These parameters are then used to generate the audio through modal synthesis.

Digital waveguides

Digital waveguide models which simulate travelling waves, are based on the discrete modelling of the wave propagation [10].

Waveguide techniques arises from the D'Alembert's solution of the ideal wave equation:

$$d^2y/dx^2 = (1/c^2) d^2y/dt^2 \tag{2.4}$$

D'Alembert stated that it can be interpreted as two travelling waves, one traveling left and other traveling right, which move at the rate c of the speed of sound in that medium:

$$y(x,t) = y_L(t + x/c) + y_R(t - x/c) \tag{2.5}$$

Therefore, the solution of the wave equation can be seen as a sum of travelling waves, which can be represented as delay lines. Digital filters are used to simulate traveling-wave attenuation [6; 10]

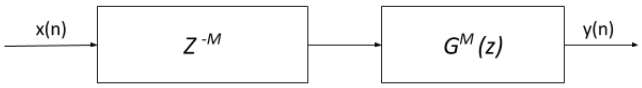


Figure 2.2: Simple simulation of a travelling wave. The filter $G(z)$ represents the propagation losses

Digital waveguides have been widely used for vibrating strings and wind sounds.

Chapter 3

Object detection

3.1 Introduction

Object detection systems are able to find objects in videos and images that belong to a certain class such as cars, dogs or faces. For humans, this task is performed effortlessly but for computers, which see the images as an array of pixel values, the process is more challenging.

Detecting and recognizing objects in an image is a challenge that has been extensively addressed in the computer vision field. There exist a large number of approaches that have been developed and which use multiple different techniques. These techniques include classification methods as support vector machines (SVM) or neural networks. Object detection systems use datasets of annotated images with labeled objects and extract characteristics of them like geometric forms or colours using a particular algorithm. A very broadly summary of these techniques is presented below.

3.2 Techniques

3.2.1 Feature-based techniques

Feature-based methods [11; 12], which are based on the extraction of features from the image, establish correspondences with a database containing the features extracted from the objects of interest. These characteristics can be borders, corners or even colours. These methods, intended for image recognition, have been commonly adapted and extended to video.

3.2.2 Template matching

Template matching is a technique in which object recognition is achieved by finding parts of an image that match a stored template [13]. Two main techniques are used in *fixed template matching*: image subtraction, where the goal is to minimize

the distance function between the aligned template and different parts of the image, and correlation, where the similarity between the segments of the input image and the template is measured by cross-correlation or normalized cross-correlation. *Deformable template matching* process are appropriate for non-rigid bodies that vary due to deformations of the object or because of different transformations relative to the camera, like rotation and scaling.

3.2.3 Motion detection

Background subtraction

A simple and widely used technique for object segmentation in static scenes [14] is background subtraction. It compares the input image with a reference background images that doesn't contain any objects of interest and threshold the result to detect new objects. If there is a significant difference between the input image and the background image, then it means that there is an object of interest in that area. Although background subtraction techniques can perform well for some applications, such as video surveillance, they are sensitive to sudden illumination changes or changes in the background (for example, if there was a piece of furniture that then is moved or removed).

Optical flow methods

Optical flow methods are commonly used techniques based on the distribution of apparent motion of objects that arise from the relative motion between the objects and the viewer [15]. The common hypothesis in measuring image motion is that pixel intensities are translated from one frame to the next. The optical flow constrain equation is defined as

$$\nabla I v + I_t = 0 \tag{3.1}$$

where $\nabla I = (I_x, I_y)$ and I_t are the spatial and temporal partial derivatives of the image I and $v = (u, v)$ is the velocity of the image. This equation with two unknowns cannot be solved as such (*aperture problem*) and therefore additional constraints must be introduced to determine the flow. However this won't be developed any further as it exceeds the scope of this thesis.

3.2.4 Convolutional neural networks

Introduction

Artificial neural networks (ANNs) are connectionism models (movement emerged in the context of cognitive science) based on a collection of connected units called neurons. The area of ANNs was inspired by the way biological neural systems process information.

The basic unit of the brain is a neuron. Each neuron receives inputs that have associated weights and computes an output by applying a function, called *activation function* to the weighted sum of inputs. Commonly used activation functions are [16]:

- Sigmoid. The sigmoid function squashes the input into the range [0, 1]. This function saturates when its arguments are large negative or large positive numbers. Large negative numbers become 0 and large positive numbers become 1. Therefore this function becomes very insensitive to small changes in its input.

$$\sigma(x) = 1/(1 + e^{-x}) \tag{3.2}$$

- Tahn. Tahn function takes its input and squashes it to the range between -1 and 1. It is similar to the sigmoid function but in this case, the output is zero-centered. It has the mathematical form

$$tanh(x) = 2\sigma(2x) - 1 \tag{3.3}$$

- ReLu. This activation function, popular in modern neural networks, thresholds the input at zero. It is defined as

$$f(x) = max(0, x) \tag{3.4}$$

Feedforward neural networks consist of multiple neurons arranged in layers. These networks are called feedforward because there are no feedback connections. When these networks include feedback connections, they are called recurrent neural networks. Convolutional networks are a specific kind of neural networks that use convolution in at least one of their layers [17]. The network is formed by three types of layers: the input layer, which receives the inputs; the output layer, which is the last fully-connected layer that produces the output of the network; and the hidden layers that are located in between these two.

Neural networks can be be formed by numerous hidden layers (*deep learning*). The overall length of the layers gives the depth of the model.

The input to a convNet is an array of pixel values. These input images are convoluted with a filter to obtain *feature maps*. The succeeding hidden layers can be seen as individual feature detectors which recognize more sophisticated patters as it is propagated through the network. The pooling (downsampling) layers further reduce the size of the representation to minimize the amount of parameters and computation of the network. In this way, the original image is transformed layer by layer from the pixel values until obtaining a one-dimensional vector, which represents the final classes scores.

In order to train the network some feedback about its performance must be provided. This is done by the *loss function*, which measures the accuracy of the prediction considering the ground truth. The objective is that the predicted label is

the same as the training label or in other words, to minimize the loss of the network by changing the weights between the different neurons. A popular method randomly initializes the weights and tunes them iteratively by moving on the direction given by the descendent gradient of the loss function. The *learning rate* is a parameter that can be adjusted to find a minimum in the function and that determines the size of the step in every iteration.

Applications

Besides traditional computer vision techniques [18; 11], convolutional neural networks (ConvNets) have been widely applied, either for car plates [19] or action recognition [20], toys recognition[21] or scenes and objects classification in photographs [22; 23]. With the advent of large datasets of images, ConvNets have been able to move forward such as with the 1000-classes ImageNet dataset [24], which contains millions of labeled images and that is used in the ImageNet Challenge [25]. One particular case is the versatile network "Alexnet" [22] that has been applied to a diversity of computer vision applications such as object detection, video classification [26] and segmentation [27].

3.3 Generating sound from silent videos

Recent work has used neural networks to synthesize sound from silent videos of a person using a drumstick to hit and scratch different objects [28], or to recognize objects and scenes from the sound of videos [29] Likewise, in [30], Daves et al. recover audio from high-speed videos of objects that stir in response to the sound. They extract those vibrations and recover the original sound that produced them. Other methods automatically generate contact sounds by using physical parameters obtained from animated simulations [2]. Similarly, in [31], the algorithm generates a soundtrack for an animation by using other animations soundtracks.

Chapter 4

Implementation

A system that helps the sound designer in the process of adding audio effects to videos was implemented. An overview of the algorithm can be seen in the graph below (Figure 4.1).

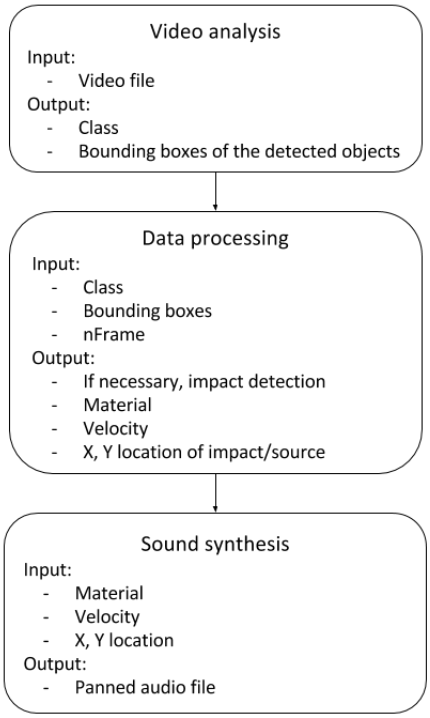


Figure 4.1: Flow chart of the implemented method

The tool can be divided in three modules: *video analysis, data processing and sound synthesis*. The input of this system is a silent video file recorded with a static camera and the output is an audio file that matches the action performed in the

movie. The video is analyzed with the help of a neural network, which detects the objects present in the movie, i.e., it labels the the items in every frame, and then this data is sent to the *data processing* module.

After receiving this information, the velocity of a moving object can be computed and, if there is more than one object in the scene, impacts are detected.

This information is sent to the sound synthesis module which generates the specific audio associated to the performed action and material.

4.1 Video analysis

4.1.1 Object detection

Preparation of the dataset

The images used to train the network were pulled out from the ImageNet database. ImageNet is a large image database comprised of more than 14 million of images and intended for object classification and detection. The ImageNet [24] labels are taken from a language database called WordNet [32], which organize words (nouns, verbs, adjectives and adverbs) into sets of synonyms called synsets (synonym sets).

These synsets are interconnected according to their semantic and lexical relations. For example, in WordNet oak and birch are hyponyms or subordinates of tree, which is a type of plant, which is an organism, etc. ImageNet database is arranged according to the hierarchy of nouns of WordNet and each node is constituted by hundreds or thousands of images. The average number of images is over five hundred per node.

However, despite the large amount of synsets, there are none intended for material detection. Therefore, the dataset used was created by combining images of different classes taken from ImageNet and additional pictures taken by the author. A more detailed description can be found in the appendix A.

The downloaded synsets contain thousands of pictures. However, just a little more than a hundred of labeled bounding boxes are available in the dataset. Therefore, the images that did not have their corresponding labeled data were discarded. The remaining images were grouped in different classes and labeled as *metal, wood, glass or motor*.

Finally, the bounding boxes files were converted from PASCAL (Visual Object Classes) VOC [33] format (.xml file with the metadata) to Darknet format (text file containing only the information about the class and the bounding boxes).

In addition to the images downloaded from ImageNet, more images were taken, hand-labeled [1] and added to the dataset.

Once we had all the images and their corresponding annotations, the train and test files were generated.

[1] Using the software provided by https://github.com/puzzledqs/BBox-Label-Tool

Data augmentation To increase the size of the dataset and make the model more robust, before creating the train and test files, all the pictures were processed. Gaussian noise was added to the pictures and their brightness was reduced by a factor of 2 and 3, and increased by a factor of 1.5.

Figure 4.2: Example of the modified images. From upper left to bottom right: original photo, photo with brightness increased, photo with brightness reduced by a factor of 2, photo with brightness reduced by a factor of 3, with gaussian noise of mean = 0.05 and gaussian noise with mean = 0.2

In this way the number of images for all the classes was increased. In total, 23293 images were used for training and 5950 for testing.

YOLO object detection system

YOLO (You only look once) [34] is a open source state-of-the-art object detector which is able to process video in real time. It is implemented in a neural network framework, Darknet [35], that was developed by Joseph Redmon. This detector is used to obtain the bounding boxes and the classes of the objects in the video.

YOLO, which is inspired by the GoogleNet model for image classification, reshape object detection as a regression problem [34] and split the input image into an SxS grid, predicting SxSx(B*5+C) values. Each cell of the grid predicts C conditional class probabilities and B bounding boxes and their corresponding confidence scores,

which represent how confident the system is that the box contains an object and how accurate.

The generated bounding box is an area with sides parallel to the X,Y coordinate axes and that defines the rough size of an object.

The network predicts 5 coordinates for each bounding box: t_x, t_y, t_h, t_w and t_o. The center coordinates of the bounding boxes are (t_x, t_y) and (t_w, t_h) are the width and high. All these values are in fractions of image size. So, for instance, $(t_x, t_y) = (0.5, 0.5)$ is the center of the image and if $t_w = 1, t_h = 1$, the bounding box is the size of the input image. To determine the coordinates in pixels, a conversion must be done:

$$\begin{aligned} left &= (t_x - t_w/2) * w \\ right &= (t_x + t_w/2) * w \\ top &= (t_y - t_h/2) * h \\ bot &= (t_y + t_h/2) * h \\ x &= t_x w \\ y &= t_y h \end{aligned} \tag{4.1}$$

where w and h are the width and high of the original image.

The network was trained with the customized dataset of images, with a learning rate of 10^-3, a momentum of 0.9 and the pretrained weights provided by the author of the model [35]. The training was done in a computer with Ubuntu 16.04 operating system using a Titan X graphics card and it took 3 days to complete 80000 epochs. This was empirically tested 6 times until a good result was found.

The output classifies the detected objects into 4 different classes: wood, glass, metal and engine.

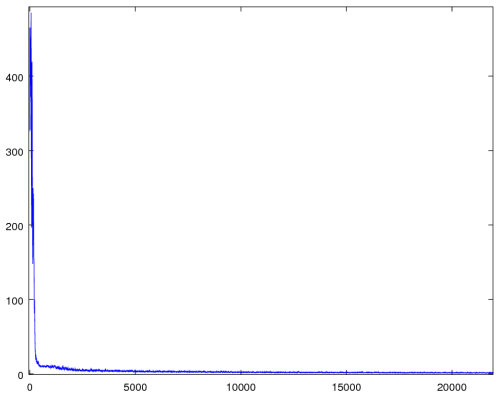


Figure 4.3: Plot of the loss value. Y-axis is the error and X-axis the number of epochs.

4.1.2 Postprocessing of the file

In this step, the file containing the bounding boxes and classes is modified in order to fix possible errors and to compute the velocity which will be used afterwards to generate the sound. The detector does not perform perfectly so sometimes the objects disappeared suddenly in some frames. This problem arises mainly while performing fast movements, like in the case of strong knocks, where the moving object becomes blurry and therefore it is more difficult to detect. This problem is solved by adding the missing object to the file in those frames where it has disappeared. The bounding boxes' coordinates of the missing objects are computed by calculating the mean value of the previous and consecutive frames. This process is applied both for videos that contain one or two elements.

The text file is processed to check how many objects were detected in the whole video. The system has a constraint of a maximum of only two objects to keep a simplicity in the model. If two objects were found in the video, then an impact detection is conducted.

Impact detection

A second process of the file is done if there are two objects detected in the video. This is performed in order to detect if an impact has occurred between both objects. The first step detects collision between the two items. This is done by verifying that the two bounding boxes are overlapped or adjoining. However, as stated above, the bounding boxes are parallel to the axes and this is problematic if the objects are not facing the desired direction or if they have a shape that differs a rectangle.

Therefore it is not enough just to verify that the bounding boxes are overlapped to guarantee if two objects have touched.

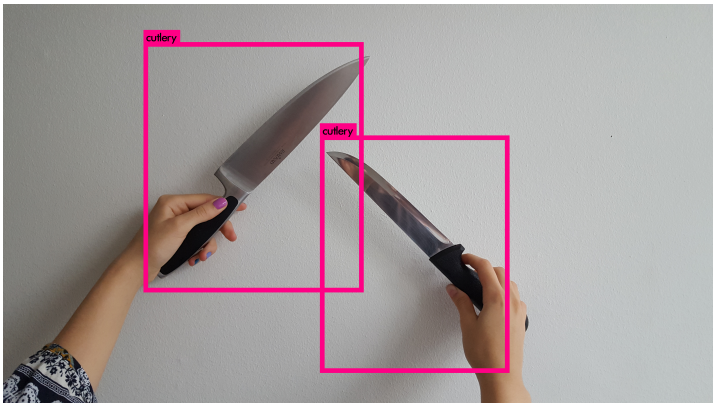


Figure 4.4: Detected bounding boxes of two knifes. The bounding boxes are overlapped but the real objects are not touching.

To ensure that one object has hit the other one, the velocities must be additionally checked.

The velocities of the objects are estimated by finding the displacement of the respective centroids of the bounding boxes between two consecutive frames.

$$v = \frac{x_1 - x_0}{\Delta t} = \frac{\Delta x}{\Delta t} \tag{4.2}$$

After computing the means of the velocity of both objects, the system can discern which object is the one that is moving and which one is remaining still. This is necessary to compute the relative velocity of the moving object which is used to detect the impact. This is also needed to determine the material involved in the action. The object that is struck is considered as the one that wants to be sonified. Thus, the sound produced will have the characteristics of the material of the object that has been hit.

Once the velocity vector is obtained, peaks in the array are detected. When the velocity reaches its maximum and then drops or changes its sign (i.e. the object bounces to the opposite direction), there is the possibility that the object has hit something.

If both conditions have been fulfilled, then the system determines that an impact has occurred. The information containing the frame where the impact has been detected as well as the location of it, is sent to the sound synthesis module. The location of the impact corresponds to the point where the two bounding boxes are overlapping or touching.

The velocity is given in pixels per frame. Before sending this information to the next module, a conversion to meters per second is done by assuming that the space from side to side at the objects distance is 1 meter. This is a very rough approximation that can be easily changed afterwards if the result does not satisfy the sound designer.

After the file is processed, it can be sent to the sound synthesis/spatialization module.

4.2 Sound synthesis

4.2.1 Impact sounds

The last step of the system makes use of the well know modal synthesis technique to model the sounds generated by rigid bodies.

Modal synthesis can be seen as a connection of second order resonant filters [36]:

$$y(n) = 2R\cos(\theta)y(n-1) - R^2 y(n-2) + aF(n-1) \tag{4.3}$$

$$R = e^{-d/fs} \tag{4.4}$$

$$\theta = \omega/fs \tag{4.5}$$

where d is the decay rate, fs is the sampling frequency and $\omega = 2\pi f$ is the frequency of the mode. As stated above, the resulting sound depends on many

factors as the shape and dimension of the object, the impact velocity or the location of the collision. The literature has shown that the perception of the material in impact sounds is mainly based on the frequency-dependent damping of the spectral components (equivalently, the sound decay) and the spectral content of the sound [37].

The parameters associated to the sound of each material (wood, glass and metal) can be extracted experimentally by analyzing real recordings and fitting the model parameters to the recorded sound. This was done for glass, metal and wooden sounds.

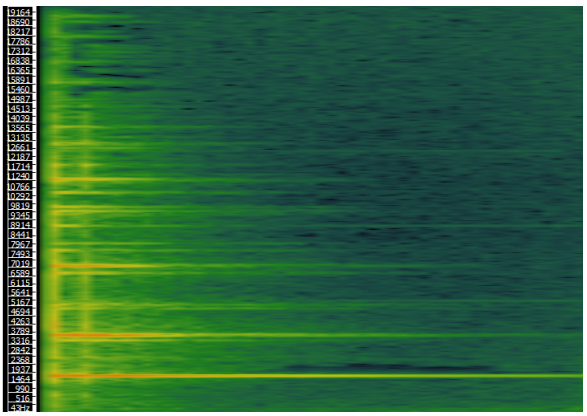


Figure 4.5: Spectrogram of the glass sound. The main modes can be easily appreciated

Wooden sounds, characterized by a low pitch and rapid decay, have larger decay rates than metal and glass sounds that are characterized by long decay times.

Several signals were tested to generate the excitation of the system: impulses, bursts of noise and the model proposed in [2] $1 - \cos(2\pi t/\tau)$, where $0 \leq t \leq \tau$ and τ is the total duration of the contact. However the sounds generated by these models weren't plausible enough, so it was decided to use the residual from the recordings of the different struck objects by using inverse filtering of the main modes.

Engine sound

The sound for the car and motorbike was generated using the Andy Farnell's engine model [38]. It is based in a four-stroke engine which produces its characteristic noise due to the gas expelled at high pressures by the pistons and that resonates inside the exhaust system. He considers some sound sources [38]:

- Explosive pulses radiated directly from the engine
- Pulses that are coupled through the vehicle body
- Radiation from the exhaust pipe surface
- Pulses from the tailpipe

- Other sounds like tyres, fanbelt etc.

The source of energy is generated by a sawtooth wave split into various subphases. The engine, exhaust, and body can be seen as a series and parallel network of excitations, modeled by a combination of delays, filters, phase splitting and wrapping. A description of the model can be seen in Figure 4.6

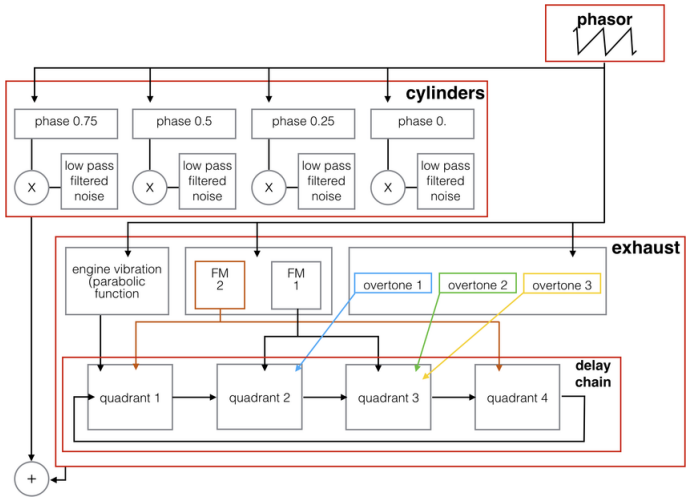


Figure 4.6: Block diagram of the Andy Farnell's engine model. Taken from [39]

4.2.2 Sound panning

After the mono sound is generated, it is panned according to the relative position of the source in the video.

The generated mono sound as well as the information containing the coordinates of the object/impact location that are present in the video image are fed into this last function which locate it and produces a stereo output sound.

4.3 Limitations

- Perspective. Two bounding boxes can be overlapped without being touched. For example, if there are two objects, one behind the other, and one of it performs a sharp movement, an impact will be wrongfully detected.
- Bounding boxes. A rectangle is a very rough shape that barely indicates the location of the detected object. This leads to problems in the impact detection as well as for getting a more accurate sound that matches the struck object since it is not possible to know accurately in which part has been hit.
- Both objects (the one that strikes and the one that is hit) need to be detected by the network. So for example, if the moving object was a plastic stick with

an uncommon shape, the model wouldn't detect it and therefore no impacts would be searched.

Chapter 5

Evaluation

To asses the quality of the generated sounds and how well they match real objects, a perceptual evaluation was performed.

5.1 Study design and stimuli

Two cases were studied: sounds without any video and videos sonified using different methods. The test was based in two main hypothesis:

- Audio sounds better if there is a visual input. The quality of the sound is rated higher when there is a video that accompany the audio.
- The developed system has the same performance as a manually matched work. The sounds automatically generated by the model fits the videos as good as the recorded and manually modified audios.

For the first case, in which no visual input was given, two examples of glass, wood and metal sounds were evaluated: synthesized sounds and recorded audio clips (downloaded from `freesound.org`, a collaborative sound database).

In the second part of the test where the subjects watched a short video, glass, wood and metal sounds were considered as well as a motor sound. To obtain the corresponding audio, three methods were used:

- Sounds generated by modal synthesis. To obtain the characteristic sounds for each material some recordings were done and the main modes were extracted from the corresponding spectrograms.
- Modal synthesis sounds manually matched. In other words, the same synthesis method was used but the timing and the parameters of velocity, location and sound volume were manually changed to match the video.
- Recorded sounds. The sounds were downloaded from `freesound.org`, the synchronization with the video was done manually and the amplitude was changed according to the velocity of impact.

In addition to these sounds, an anchor, consisting in the low-pass filtered version of the excitation sounds for each material, was added. The panning and the synchronization were also modified to be wrong.

For motor sounds only two cases were studied:

- Physical modelling based on the Andy Farnell's synthesis model [38]. The parameter of the engine's velocity and panning were automatically modified according to the apparent velocity extracted from the moving objects in the video.
- Recorded sounds. The sounds were downloaded from `freesound.org` as it was done for the previous examples. The panning and volume were manually adjusted to match the video.

A total of 4 videos were shown (see Appendix B). They were recorded with the rear camera of a Samsung Galaxy S6 in a typical kitchen scenario. In particular, the videos recorded were:

- A knife hitting a glass jar
- A knife hitting a metallic pot
- A glass jar hitting a wooden table
- A clip of a movie in which a man rides a motorcycle

Figure 5.1: Frames extracted from the videos of small impacts

5.1.1 Procedure

The participants filled in a questionnaire (see Appendix B) that contained the videos to be studied and the links to the audio files.

Before starting the evaluation, the subjects were asked about their experience in sound design and the hours that they spent watching movies or playing videogames.

In the first part of the test only audio stimuli was given to the subjects. They listened some examples of the sounds and were asked to rate the quality and choose the apparent material of the sound that was played. In the second part of the test the subjects watched the videos sonified by the different models. They were

asked about the quality of the sound considering the objects involved in the scene and the matching of the video with the audio. Finally they were asked about their preferences.

5.2 Results and discussion

A total of 15 participants aged between 24 and 60 volunteered in the evaluation. Only 4 of them spent more than 5 hours per week watching movies/series and 2 playing videogames. This number of hours is not extraordinary high so this information wasn't take into consideration for the analysis. The subjects that rated their sound design experience with more than 5 in a scale of 10 were considered expert listeners.

5.2.1 Evaluation of audio stimuli

A *t-test* was conducted in the normal-distributed data. The results for the first part of the test showed that the listeners didn't perceived significant differences between the recorded and synthesized sounds of metal ($p = 0.91, \alpha = 0.05$) and glass ($p = 0.16$).

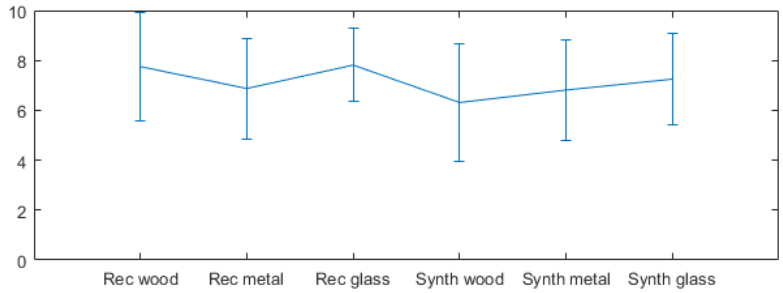


Figure 5.2: Means and standard deviations of the measured qualities. *Rec* stands for recorded sounds and *Synth* for synthesized sounds.

However, the subjects rated the synthesized wooden sounds with a lower quality than the recorded one. A *t-test* found significant differences ($p = 0.03, \alpha = 0.05$) between the recorded and the synthesized sounds of wood at 5% significance level. To be sure about these results, the Bonferroni correction was applied ($\alpha' = 0.017$), and no significant differences were found in any of the three cases.

A comparison between the results obtained in the first part of the test (only audio input) and the second one (audiovisual input) revealed that -although the means of the sound quality of the generated and recorded sounds were slightly higher in the cases where a visual input was given- there are not significant differences if a visual cue is added. After computing a Two One-Sided Test (TOST) test for equivalence, it was proved that the results were equivalent only for the synthesized sound of glass ($p = 0.009 \alpha = 0.05, diff.mean = 0.7$).

5.2.2 Evaluation of audiovisual stimuli

The second part of the test suggest that people still prefer recorded sounds than synthesized ones. Almost all of the subjects -12 to be precise- preferred the recorded sound for the video of the table. In the case of the engine the number of volunteers that preferred the recording was 10, for the video of the pot there were 9 and 8 for the glass.

In the question regarding the matching of the video and audio, the best results were found for the case of glass sounds. An ANOVA test showed that there were statistical differences between groups for the three cases at a significant level of 0.05 (wood, $p = 0.910^{-}5$, metal $p = 0.310^{-}4$, and glass, $p = 0.210^{-}3$), which was expected since there is an anchor signal. Significant differences were also found between the recorded and automatically generated sound for wood ($p = 0.0043$) and metal ($p = 0.110^{-}7$) but not for glass ($p = 0.1$). A TOST test revealed that the recorded and automatically generated sounds for glass can be considered equivalent ($p = 0.03, hypoth.mean = 0.7, \alpha = 0.05$).

No significant differences were found between the recorded and the generated sounds -those manually modified to match the video- and the expert listeners even preferred the synthesized one sometimes.

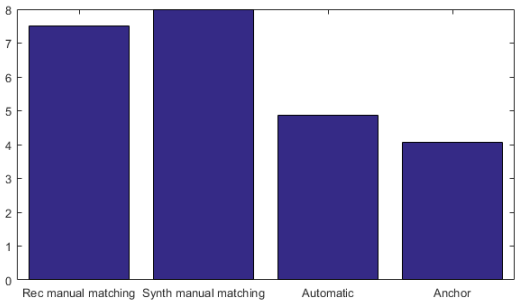


Figure 5.3: Means of the answers to the question *Does the sound match the video?* in the video of the glass jar (expert listeners)

Although on average the expert listeners rated the matching of synthesized manually-matched sound higher and a TOST test revealed that the results of the recorded and the manually-matched generated sounds can be considered equivalent ($p = 0.03, \alpha = 0.05, mean = 0.7$).

Similar results were found in the case of wooden sounds. Although the recorded sound was preferred, no significant differences were found with the video that has manually-matched synthesized audio. A TOST test found that they can be seen as equivalent.

In the case of the video of the motorcycle, the subjects rated the adequacy of the sample-based video with a higher score than the synthesized one. A t-test confirmed that there were significant differences ($p = 0.004, \alpha = 0.05$) between the two videos.

Therefore, our second hypothesis could not be proved neither since the automatically generated and recorded sounds were considered equivalent in the case of the glass.

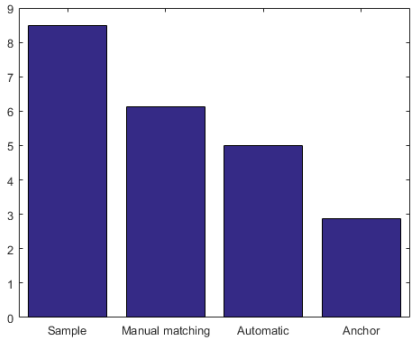


Figure 5.4: Matching rates for the video of the glass jar (expert listeners)

The worst results were found in the videos with the metallic sound. This may be possible due to the shape of the object. The sound was marked as metallic in the first part of the test but it didn't fit very well with the object displayed in the video. The generated sound had bright modes with long decays and the recorded sound was more noisy. The subjects commented that the timbre of the synthesized sound didn't match the video because of the location of the struck. Therefore one can consider that the synthesized sound could have worked well if the knife had hit the body of the pot, where a clear metallic sound is expected. However the pot was struck on its edge, which entails a more noisy sound.

These results indicate that the synthesized sounds are good enough to be added to real videos but that the parameter estimation must be improved since the subjects stated in some cases that the synchronization of the hit was not perfect nor the suitability with the location of the hit.

Chapter 6

Conclusion

A system composed by a object detector, impact detector and sound modeler was proposed. This system would speed up and facilitate the work of sound designers when they sonify a video, which can be a very tedious labour. The system would automatically generate the audio of simple actions. In this way the worker could focus on more important sounds and put less effort in matching the sound of these simple actions.

The results of the test show that the sounds generated by the system were acceptable to be added to a realistic video. However, the system performance is not perfect. The main problems were due to the lack of accuracy of the object detection stage. Bounding boxes are not exact enough to determine the real position of an object and a better model should be used to get better results. In addition, since the video provides a 2D image, there is no distance involved.

There is still much work to do. The video analysis must be improved to get a more accurate model of the objects or materials involved in the actions. In addition, other networks could be more advantageous as one that performs object segmentation - partition the video image into regions- as well. By performing segmentation we could have a better model of the image and assume its shape which lead a better impact detection and timbre of the sound. With a better detection of the objects, other actions could be identified, such scrapping or rubbing and other methods intended for sound synthesis, like digital waveguide synthesis could be used.

If the system were improved and implemented in real time it could result in artistic applications or games as well. With the use of a webcam, the system could detect some objects or actions and generate a response, such a musical or cartoon sound.

Therefore, one can see a promising future for systems that combine the best tools of different fields.

Bibliography

[1] Wikipedia, "Vitaphone — Wikipedia, the free encyclopedia," 2004. [Online]. Available: https://en.wikipedia.org/wiki/Vitaphone

[2] K. Van Den Doel, P. G. Kry, and D. K. Pai, "Foleyautomatic: physically-based sound effects for interactive simulation and animation," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, 2001, pp. 537–544.

[3] D. B. Lloyd, N. Raghuvanshi, and N. K. Govindaraju, "Sound synthesis for impact sounds in video games," in *Symposium on Interactive 3D Graphics and Games*. ACM, 2011, pp. PAGE–7.

[4] C. Zheng and D. L. James, "Rigid-body fracture sound with precomputed soundbanks," *ACM Transactions on Graphics (TOG)*, vol. 29, no. 4, p. 69, 2010.

[5] R. A. Garcia, "Automatic generation of sound synthesis techniques," Ph.D. dissertation, Citeseer, 2001.

[6] J. O. Smith, *Spectral Audio Signal Processing*. `http://ccrma.stanford.edu/~jos/sasp/`, online book, 2011 edition.

[7] F. Dunn, W. Hartmann, D. Campbell, and N. Fletcher, *Springer handbook of acoustics*. Springer, 2015.

[8] G. De Poli, "A tutorial on digital sound synthesis techniques," *Computer Music Journal*, vol. 7, no. 4, pp. 8–26, 1983.

[9] Z. Ren, H. Yeh, and M. C. Lin, "Example-guided physically based modal sound synthesis," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 1, p. 1, 2013.

[10] S. Serafin, "The sound of friction: real-time models, playability and musical applications," Ph.D. dissertation, stanford university, 2004.

[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.

[12] M. Brown, D. G. Lowe *et al.*, "Recognising panoramas." in *ICCV*, vol. 3, 2003, p. 1218.

[13] R. Brunelli, *Template matching techniques in computer vision: theory and practice*. John Wiley & Sons, 2009.

[14] A. M. McIvor, "Background subtraction techniques," *Proc. of Image and Vision Computing*, vol. 4, pp. 3099–3104, 2000.

[15] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM computing surveys (CSUR)*, vol. 27, no. 3, pp. 433–466, 1995.

[16] "Cnvolutional neural networks for visual recognition course," https://cs231n.github.io/neural-networks-1, [Online; accessed May 2017].

[17] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[18] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2. Ieee, 1999, pp. 1150–1157.

[19] R. Parisi, E. Di Claudio, G. Lucarelli, and G. Orlandi, "Car plate recognition by neural networks and image processing," in *Circuits and Systems, 1998. ISCAS'98. Proceedings of the 1998 IEEE International Symposium on*, vol. 3. IEEE, 1998, pp. 195–198.

[20] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[21] Y. LeCun, F. J. Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2. IEEE, 2004, pp. II–104.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[26] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[28] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2405–2413.

[29] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems*, 2016, pp. 892–900.

[30] A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, and W. T. Freeman, "The visual microphone: passive recovery of sound from video," 2014.

[31] M. Cardle, S. Brooks, Z. Bar-Joseph, and P. Robinson, "Sound-by-numbers: motion-driven sound synthesis," in *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*. Eurographics Association, 2003, pp. 349–356.

[32] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database," *International journal of lexicography*, vol. 3, no. 4, pp. 235–244, 1990.

[33] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[34] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *CoRR*, vol. abs/1506.02640, 2015. [Online]. Available: http://arxiv.org/abs/1506.02640

[35] J. Redmon, "Darknet: Open source neural networks in c," http://pjreddie.com/darknet/, 2013–2016.

[36] K. Van Den Doel and D. K. Pai, "Modal synthesis for vibrating objects," *Audio Anectodes. AK Peter, Natick, MA*, pp. 1–8, 2003.

[37] D. Rocchesso and F. Fontana, *The sounding object*. Mondo estremo, 2003.

[38] A. Farnell, *Designing sound*. Mit Press, 2010.

[39] S. Baldan, H. Lachambre, S. Delle Monache, and P. Boussard, "Physically informed car engine sound synthesis for virtual and augmented environments," in *Sonic Interactions for Virtual Environments (SIVE), 2015 IEEE 2nd VR Workshop on*. IEEE, 2015, pp. 1–6.

Appendix A

Dataset details

The dataset used to train the network was a combination of the following synsets taken from Imagenet:

- Glass
  - Mason jar. Wnid: n03725600
  - Wine bottle. Wnid: n04591713
  - Beer bottle. Wnid: n02823428
  - Water glass. Wnid: n04559910
- Wood
  - Dinning table, board. Wnid: n03201208
  - Dining-room table. Wnid: n03201035
  - conference table, council table, council board. Wnid: n03090000
- Metal
  - Pan, cooking pan. Wnid: n03880531
  - Frying pan, frypan, skillet. Wnid: n03400231
  - Caldron, cauldron. Wnid: n02939185
  - Knife. Wnid: n03624134
  - Knife. Wnid: n03623556
  - Spoon. Wnid: n04284002
- Motor
  - Motorcycle. Wnid: n03790512

Wnid stands for *WordNet ID*

Appendix B

Questionnaire

Gender *

- Male
- Female

Age *

How much time do you spend on average watching movies/series? *

- < 2 hours per week
- 2-5 hours per week
- 5-10 hours per week
- > 15 hours per week

How much time do you spend on average playing videogames? *

- < 2 hours per week
- 2-5 hours per week
- 5-10 hours per week
- > 15 hours per week

Sound design experience *

	1	2	3	4	5	6	7	8	9	10	
None	○	○	○	○	○	○	○	○	○	○	Very experienced

Part 1

Now you can start with the first part of the test!

1-Listen to the sound and answer the following questions

https://drive.google.com/file/d/0B1PPtHEA11xGd1R5ZzkyX1o4STQ/view?usp=sharing

Which material is involved in the action? *

- [ ] Wood
- [ ] Metal
- [ ] Glass
- [ ] Otro: _____

Please rate the quality of the sound *

	1	2	3	4	5	6	7	8	9	10	
Very bad	○	○	○	○	○	○	○	○	○	○	Excellent

Comments (optional)

Tu respuesta

2-Listen to the sound and answer the following questions

https://drive.google.com/open?id=0B1PPtHEA11xGVXJlelo5MXBzZE0

Which material is involved in the action? *

- [ ] Wood
- [ ] Metal
- [ ] Glass
- [ ] Otro: _____

Please rate the quality of the sound *

	1	2	3	4	5	6	7	8	9	10	
Very bad	○	○	○	○	○	○	○	○	○	○	Excellent

Comments (optional)

Tu respuesta

3-Listen to the sound and answer the following questions

https://drive.google.com/open?id=0B1PPtHEA11xGVWo4TWphOXlNUFk

Which material is involved in the action? *

- [ ] Wood
- [ ] Metal
- [ ] Glass

Please rate the quality of the sound *

	1	2	3	4	5	6	7	8	9	10	
Very bad	○	○	○	○	○	○	○	○	○	○	Excellent

Comments

Tu respuesta

4-Listen to the sound and answer the following questions

https://drive.google.com/open?id=0B1PPtHEA11xGMjMzVXpIWEdHdVU

Which material is involved in the action? *

- ☐ Wood
- ☐ Metal
- ☐ Glass
- ☐ Otro:

Please rate the quality of the sound *

	1	2	3	4	5	6	7	8	9	10	
Very bad	○	○	○	○	○	○	○	○	○	○	Excellent

Comments

5-Listen to the sound and answer the following questions

https://drive.google.com/open?id=0B1PPtHEA11xGYWFqYTNWMUtDUzA

Which material is involved in the action? *

- [ ] Wood
- [ ] Metal
- [ ] Glass
- [ ] Otro: _____

Please rate the quality of the sound *

	1	2	3	4	5	6	7	8	9	10	
Very bad	○	○	○	○	○	○	○	○	○	○	Excellent

Comments

Tu respuesta

6-Listen to the sound and answer the following questions

https://drive.google.com/open?id=0B1PPtHEA11xGbUFZZTU3STNZakk

Which material is involved in the action? *

- [ ] Wood
- [ ] Metal
- [ ] Glass

Please rate the quality of the sound *

	1	2	3	4	5	6	7	8	9	10	
Very bad	○	○	○	○	○	○	○	○	○	○	Excellent

Comments

7-Listen to the sound and answer the following questions

https://drive.google.com/open?id=0B1PPtHEA11xGYzllbE1mZ3duV3c

Which material is involved in the action? *

- ☐ Wood
- ☐ Metal
- ☐ Glass
- ☐ Otro: _____

Please rate the quality of the sound *

	1	2	3	4	5	6	7	8	9	10	
Very bad	○	○	○	○	○	○	○	○	○	○	Excellent

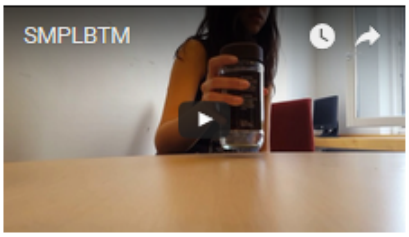
Comments

Part 2

You are almost done! This is the last part of the test. Please watch the following videos and answer their corresponding questions.
There are 4 different videos.

1 - Watch the following videos and answer the questions below

Video 1



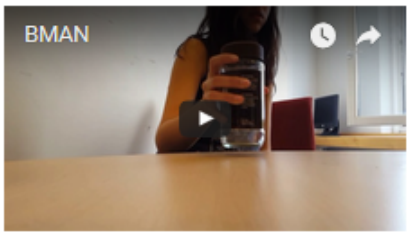
Quality of sound (regarding the objects involved) *

	1	2	3	4	5	6	7	8	9	10	
Very poor	○	○	○	○	○	○	○	○	○	○	Excelent

Does the sound match the video? (timing, volume, timbre) *

	1	2	3	4	5	6	7	8	9	10	
Not at all	○	○	○	○	○	○	○	○	○	○	Totally

Video 2



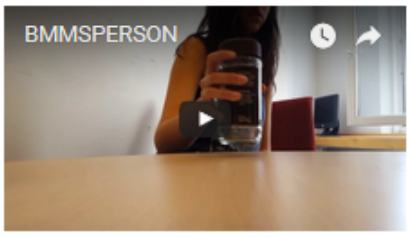
Quality of sound (regarding the objects involved) *

	1	2	3	4	5	6	7	8	9	10	
Very poor	○	○	○	○	○	○	○	○	○	○	Excelent

Does the sound match the video? (timing, volume, timbre) *

	1	2	3	4	5	6	7	8	9	10	
Not at all	○	○	○	○	○	○	○	○	○	○	Totally

Video 3



Which video do you prefer? *

- ◯ Video 1
- ◯ Video 2
- ◯ Video 3
- ◯ Video 4

Comments

2 - Whatch the following videos and answer the questions below

Video 1

Quality of sound *

	1	2	3	4	5	6	7	8	9	10	
Very poor	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Excelent

Does the sound match the video? *

	1	2	3	4	5	6	7	8	9	10	
Not at all	○	○	○	○	○	○	○	○	○	○	Totally

Video 3

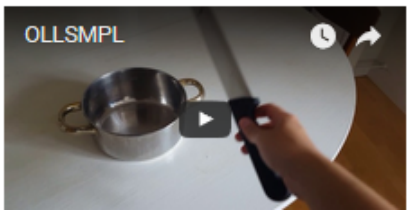
Quality of sound *

	1	2	3	4	5	6	7	8	9	10	
Very poor	○	○	○	○	○	○	○	○	○	○	Excelent

Does the sound match the video? *

	1	2	3	4	5	6	7	8	9	10	
Not at all	○	○	○	○	○	○	○	○	○	○	Totally

Video 4



Quality of sound *

	1	2	3	4	5	6	7	8	9	10	
Very poor	○	○	○	○	○	○	○	○	○	○	Excelent

Does the sound match the video? *

	1	2	3	4	5	6	7	8	9	10	
Not at all	○	○	○	○	○	○	○	○	○	○	Totally

Which video do you prefer? *

- ○ Video 1
- ○ Video 2
- ○ Video 3
- ○ Video 4

Comments

3 - Whatch the following videos and answer the questions below

Video 1

Quality of sound *

	1	2	3	4	5	6	7	8	9	10	
Very poor	○	○	○	○	○	○	○	○	○	○	Excelent

Does the sound match the video? *

	1	2	3	4	5	6	7	8	9	10	
Not at all	○	○	○	○	○	○	○	○	○	○	Totally

Video 2

Quality of sound *

	1	2	3	4	5	6	7	8	9	10	
Very poor	○	○	○	○	○	○	○	○	○	○	Excelent

Does the sound match the video? *

	1	2	3	4	5	6	7	8	9	10	
Not at all	○	○	○	○	○	○	○	○	○	○	Totally

Video 3

Quality of sound *

	1	2	3	4	5	6	7	8	9	10	
Very poor	○	○	○	○	○	○	○	○	○	○	Excelent

Does the sound match the video? *

	1	2	3	4	5	6	7	8	9	10	
Not at all	○	○	○	○	○	○	○	○	○	○	Totally

Video 4

Quality of sound *

	1	2	3	4	5	6	7	8	9	10	
Very poor	○	○	○	○	○	○	○	○	○	○	Excelent

Does the sound match the video? *

	1	2	3	4	5	6	7	8	9	10	
Not at all	○	○	○	○	○	○	○	○	○	○	Totally

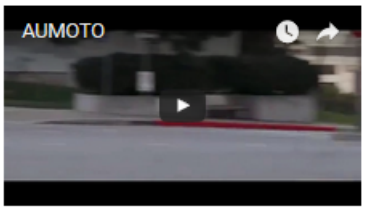
Which video do you prefer? *

- ○ Video 1
- ○ Video 2
- ○ Video 3
- ○ Video 4

Comments

4 - Whatch the following videos and answer the questions below

Video 1



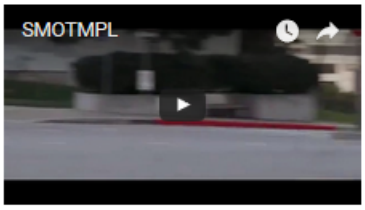
Quality of sound *

	1	2	3	4	5	6	7	8	9	10	
Very poor	○	○	○	○	○	○	○	○	○	○	Excelent

Does the sound match the video? *

	1	2	3	4	5	6	7	8	9	10	
Not at all	○	○	○	○	○	○	○	○	○	○	Totally

Video 2



Quality of sound *

	1	2	3	4	5	6	7	8	9	10	
Very poor	○	○	○	○	○	○	○	○	○	○	Excelent

Does the sound match the video? *

	1	2	3	4	5	6	7	8	9	10	
Not at all	○	○	○	○	○	○	○	○	○	○	Totally

Which video do you prefer? *

- ○ Video 1
- ○ Video 2

Comments

Links to the videos

Wood:

- Recorded sound: https://www.youtube.com/watch?v=gWoGSpHUbGA
- Synthesized manually matched sound: https://www.youtube.com/watch?v=w7CB9MocS2I
- Automatic synthesized sound: https://www.youtube.com/watch?v=RqBSLgqbfaU
- Anchor: https://www.youtube.com/watch?v=ri1bn8T_cKc

Glass:

- Recorded sound: https://www.youtube.com/watch?v=rJBSguy7ITc
- Synthesized manually matched sound: https://www.youtube.com/watch?v=QrqG0BWsiYU
- Automatic synthesized sound: https://www.youtube.com/watch?v=o-7V6qH5bJ8
- Anchor: https://www.youtube.com/watch?v=uMWqjZqJtL4

Metal:

- Recorded sound: https://www.youtube.com/watch?v=GaCU7bwGXyY
- Synthesized manually matched sound: https://www.youtube.com/watch?v=u63iVA5_Ec0
- Automatic synthesized sound: https://www.youtube.com/watch?v=6JB-uuFD-yg
- Anchor: https://www.youtube.com/watch?v=uMWqjZqJtL4

Motor:

- Recorded sound: https://www.youtube.com/watch?v=rkQa__Pp_PI
- Automatic synthesized sound: https://www.youtube.com/watch?v=F1CL7He8k3s