FOLEY SOUNDS VS. REAL SOUNDS

Stefano Trento
Conservatorio C. Pollini, Padova
trento.stefano@gmail.com

Amalia de Götzen
Sound and Music Processing Lab - SaMPL
Conservatorio C. Pollini, Padova
coordinatore@sampl-lab.org

ABSTRACT

This paper is an initial attempt to study the world of sound effects for motion pictures, also known as *Foley sounds*. Throughout several audio and audio-video tests we have compared both Foley and real sounds originated by an identical action. The main purpose was to evaluate if sound effects are always better than real sounds [1]. We found a similarity in subjects preferences between real sounds and Foley sounds, with a limited discrimination ability between them.

1. INTRODUCTION TO THE FOLEY ART

The majority of movies that are made today demonstrate such an effective and intensive use of Foley effects that their importance in animation and movie production has been widely recognized. The movie–goer will be affected by the sound so that her sonic experience will undoubtedly enhance the narrative stream of the movie. Therefore, it is appropriate to think about Foley sounds as a support for visual composition and characterization. The art of Foley grew up thanks to Jack Foley's inventiveness and open mind. He was the charismatic sound editor who invented this craft out of necessity, during the production of *Showboat*, a musical made at Universal Studios in 1929. From that moment on, with the passing of years and with the contribution of many Foley artists such as John H. Post, Ken Dufva, David Lee Fein and Robert Rutledge, this craft has become very popular in movie production [2]. The idea that Foley had was to provide a scene with sound effects by performing and adding them while the scene itself was being staged. The art of performing and creating these sounds effects consists in handling various kind of props and doing some strange movements in a special recording stage. The person who does this is called a *Foley artist*. She performs footsteps, clothes movements, props and everyday sounds both for movie, radio programs and TV shows. Essentially, she has to pay attention to what the actor is doing on screen and on account of that, she must choose the correct props to reproduce the better sound for that scene. For instance, she must observe if the actor is walking on a wooden floor or if he is beating somebody up with either his hands or any kind of prop. Her role is important because through Foley effects she can emphasize, enhance, support, replace and even create the reality of an action. The stage where Foley artists perform is called a *Foley stage* – an example of a Foley stage is shown in fig. 1. We have to think about the Foley stage as a place to design custom sound effects. Normally, it consists of two separated rooms. One is dedicated to the performance of Foley artists, while the second is used to record them. The sound technician, who records hand props, footsteps, clothes and custom effects that are needed to be performed on the Foley stage, is called a *Foley editor*.

Figure 1. Typical Foley stage, with various pits, props and monitor for sync.

Basically, a Foley stage needs an extensive technical equipment. The set up of this room is composed by: a monitor or projector, one or more pits, microphones and props. The monitor is used to help the synchronization between the action performed by the Foley artist and the action played on the screen. The pit - whose dimensions are at least 3 x 4 feet with a depth of about 6-8 inches - is used to perform the footsteps on different surfaces. For example, a pit could be filled up with coffee and another one with stones. In this way, the Foley artist can perform footsteps from different floors at the same time, without blocking the action. This is of key importance in order to be fluent and to enhance the narrative stream of the movie. Moreover, some Sound editors - known as Mixers - prefer a Foley stage with very little reflection so as to obtain an acoustically dead room. After all, a Foley effect recorded flat – that is without equalizations and prospective – is easier and quicker to process.

2. OUR FOLEY EFFECTS

2.1 Choosing Foley Effects

After a preliminary study of the Foley world, the next step was to find out which sounds – or, to be more precise, which actions – were needed for our objectives. We chose the eight different actions that follow: slapping, uncorking a bottle of wine, breaking bones, bird wings flapping, kissing, walking up the stairs, walking on summertime grass, closing a sliding door. What follows is a list which defines the Foley methods to produce each action. For the equivalent real sounds such a list is obviously not needed as it is quite simple to imagine.

- **Slapping:** this sound is created by holding a piece of raw steak with one hand and hit it with the open palm of the orther in its center. To simulate a person being slapped it is common practice to use the same method with slices of steak of different thickness depending on the part of the body being hit.
- **Uncorking a bottle of wine:** the simulation of this action is obtained by removing the piston of a big syringe previously filled with air.
- **Breaking bones:** usually, this is recreated by breaking into two halves a stick of celery in front of a microphone.
- **Bird wings flapping:** achieved by quickly wiggling a pair of leather gloves in front of a microphone.
- **Kissing:** this is done by wetting one's lips and then kissing the most hair–less part of one's forearm making sloppy kissing sounds.
- **Walking up the stairs:** there are lots of tricks [1] to perform this Foley effect. We chose the simplest and easiest method. Sitting on a chair wearing noisy sneakers a ceramics tiles surface must be hit in various ways and with different intensities.
- **Walking on summertime grass:** walking on or hitting with one's hands a 14 audiotape balled up. This Foley effect is shown in fig. 2.
- **Closing a sliding door:** this is achieved by making a roller skate slide on a piece of wood whose height is about 4 feet.

2.2 Recording Foley and real sounds

We selected our equipment according to the typical recording studio and technical equipment used by the Foley artist. We needed a low reverberant room as a Foley stage, and for this reason we ended up recording the majority of Foley and real sounds in *SaMPL*'s silent cabin - which has only 0.12 seconds of reverb time at a frequency of 1600 Hz. Some real sounds (such as: walking on summertime grass, walking up the stairs and closing a sliding door) needed a field recording session. The microphone used for the silent cabin recording session was a Sennheiser MKH 8020 - an omnidirectional microphone with a very linear frequency response curve and high sensitivity whereas for the field recording session we used a Sennheiser MKH 8040 – a cardioid microphone. Moreover, in order to achieve the best quality of audio files, we used an Orpheus FireWire audio interface. All sounds were recorded at a 48 kHz sample rate with a 24–bit resolution in order to avoid a downsampling process during the editing of the video. All the audio files that we recorded are available for the listening or the downloading on the web site `http://freesound.org` - by searching the username "stereostereo".

[1] Foley is an art, not Science. Therefore, for each Foley effect we might have different techniques to produce it. In this research we chose the simplest and the most traditional method to create Foley effects.

Figure 2. Microphone Sennheiser MKH 8020 and the 14" audiotape balled up.

3. TESTS

3.1 Preamble

The type of test needed by the analysis process has been chosen on account of the goal that we wanted to achieve. Therefore, we restricted the main objective to the direct comparison between Foley and real sounds. There are several tests that we could use to this end, but none of them is a specific standard. For that reason, we adopted an international standard [3], which is the *MUSHRA* [2], albeit with some variations. Through this method and we organized two different tests. The first was an audio–only test whereas the second was an audio–video test. In this way we could be able to compare the data stored from each test involving thus two different sensorial modalities. Moreover we programmed two graphical interfaces that allowed us to store data automatically and to control the audio and video playback (using the Max/MSP/Jitter application). All subjects were supposed to use the same type of transducer. Therefore, the equipment used for all the tests was composed by a laptop, an USB audio Interface - the Edirol ua-101, and a pair of professional Sony Mdr-7506 headphones. Furthermore, we conditioned the audio

[2] MUSHRA is the acronym of "MUlti Stimulus test with Hidden Reference and Anchors".

for each test somewhat, sometimes compressing or editing it and sometimes adding in and out fades. After that, we chose all the video fragments for the audio- video test and mixed their soundtracks with our recorded audio files. The chosen video clips were:

- Notting Hill 1999: ©Universal Pictures
- The Protector 2005: ©Eagle Pictures
- Edward Scissorhands 1990: ©Twentieth Century-Fox Film Corporation
- A Walk in the Clouds 1995: ©Twentieth Century-Fox Film Corporation

3.2 The MUSHRA method

This method was designed by the EBU project group to give a reliable and repeatable measure of the audio quality of intermediate–quality signals. It is a "double–blind multi–stimulus" test with both hidden reference and anchors [3]. In a MUSHRA test [3] the subject judges his "preference" for one type of artifact versus many others. Basically, she has to assess the impairments on "B" compared to a known reference "A" and then to evaluate "C" ("D", "E" etc.) compared to "A", where B, C, D, E are randomly assigned to a hidden reference, a hidden anchor and to the tested objects. The assessment is given according to the five-interval Continuous Quality Scale (CQS). It is a graphic scale that has a range from 0 to 100 and which is divided into five equal intervals that are: Bad, Poor, Fair, Good and Excellent.

3.3 Participants

It is very important that each participant has some experience in listening critically to the sound sequences, in order to reach results that are more reliable than those obtained with a non–experienced listener. We recruited forty experienced volunteers, twenty for the audio test and twenty for the audio-video test. All of them undertook a test which lasted less than fifteen minutes in order to avoid stress and fatigue.

3.4 Audio-Video tests

In the audio-video test ten different types of movie action were selected, defined as follows: Walking on grass, Kiss, Broken Hand, Sliding Door, Slap, Up the Stairs, Bird Flight, Bottle Cap, Double Kiss, Head and Arm Broken. For each of these actions there were three movies with the same picture but with different sounds. As a matter of fact, in one there was a Foley sound, in another one there was a real sound and in the last one there was an anchor sound [4], for a total of thirty movies [5] and an estimated total duration [6] of ten minutes and thirty seconds. The subject was asked to evaluate how realistic was each video using the CQS scale. For all the videos that needed the anchor sound we used some audio samples downloaded from `http://freesound.org`. These sounds distributed by `http://freesound.org` are licensed through a Creative Commons license, which allows changing the sounds as we want provided we give credit to the author. A list of all the anchors used along with the action they represented follows:

- `85604__horsthorstensen__walk_mud01` Walking on the summertime grass
- `66073__joerhino__DVD_BREAKING_3` Breaking Bones
- `77534__Superex1110__Glass_Crush_7` Breaking Bones
- `26341__nannygrimshaw_London_Underground` - Closing a sliding door
- `37162__volivieri__soccer_stomp_02` Walking up the stairs
- `40161__Nonoo__flobert1_20070728` Slapping
- `64401__acclivity__SwansFlyBy` Sparrow
- `8000__cfork__cf_FX_bloibb` Uncorking a bottle of wine

For the Kissing action we used the Foley sounds filtered with a high-pass filter in order to obtain a very bright and unreal sound.

3.5 Audio tests

The Audio Test included ten different types of action which were the same as those included in the Audio-Video test. Each of them contained two hidden audio files. One was the Foley sound while the other was the real sound for a total of twenty audio files and an estimated duration of four minutes and four seconds [7]. As in the video test, the listener was asked to evaluate from 0 to 100 how realistic was each audio clip compared to the action it represented. In this test there was no anchor sound as it was not essential to our main objective.

4. ANALYSIS OF DATA

4.1 Preliminary observations

Throughout the study of the data previously collected from the tests we wanted to be able to discern whether sound

[3] Generally, the anchor signal has a bandwidth limitation of 3.5kHz and is processed with a low-pass filter. Other anchor processings may include: reducing the stereo image or adding noise.

[4] On our test, the anchor sound is quite different both the real and Foley sounds. If the listener evaluated the anchor sound as a very realistic sound, his results were discarded. The anchor helps to discriminate with sufficient accuracy the correct results.

[5] Each video had a PAL 4:3 resolution whose dimension were 720 x 576 and an linear PCM audio codec - which had a sample rate of 48Khz and a depth of 24bit.

[6] The estimate was made according to two hypotheses: the listener played each video at least twice and she usedat leasttwo secondsforeachassessment.

[7] The estimate for the audio test was made according to two hypotheses: the listener played each sound at least twice and he usedat leasttwo secondsforeachassessment.

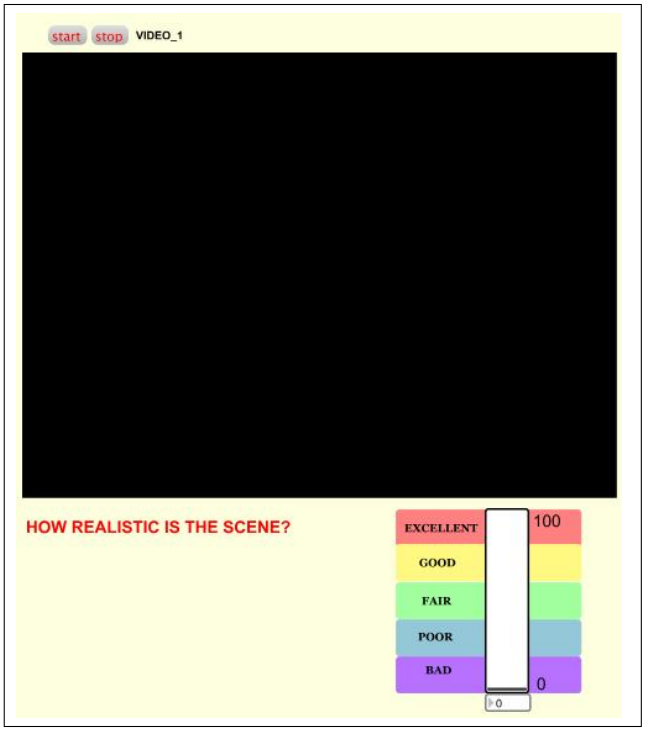

Figure 3. Interface details of the audio-video test. At the right bottom the slider of the CQS scale.

Screen action	Audio Test	Audio-Video Test
Walking on grass	Real 76.15	Foley 59.33
Kiss	Foley 54.50	Foley 71.76
Broken hand	Foley 48.85	Foley 54.25
Sliding door	Real 70.15	Foley 56.93
Slap	Foley 56.85	Real 58.42
Up the Stairs	Real 70.30	Real 52.25
Bird Flight	Real 62.45	Foley 50.57
Bottle Cap	Foley 81.65	Real 56.58
Double Kiss	Foley 67.95	Foley 76.33
Head and Arm Broken	Real 63.90	Foley 61.85

Table 1. Summary table of the highest means of each action for the audio and audio-video test.

effects are always better than real sounds. The main objective was the direct comparison between the average result of the different evaluations given to Foley and real sounds. Therefore, we have analyzed all the data with the *ANOVA* statistic method. A table with the direct comparisons between the highest means of the two tests follows.

First of all, we will analyze the summary table 1. We can observe that for the audio test the participants judged as realistic sounds five Foley sounds - the 50% of the total. On the other hand, the results for the Audio- Video test are quite different. In fact, the subjects preferred the 70% of actions with Foley sounds. This analysis was done without the ANOVA and for this reason on the next chapters we will do an accurate analysis with this method for each test.

In each test 50% of the subjects were musicians. Consequently, it is interesting to understand if the evaluations differ between musicians and non–musicians. To this end

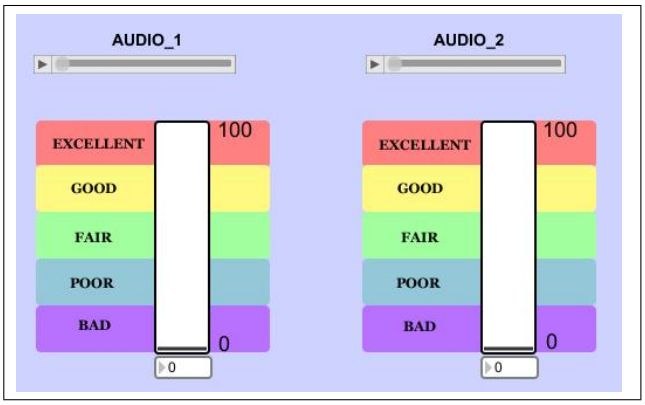

Figure 4. Interface details of the audio test. At the top there are two playbars for the audio playback.

we performed a separate ANOVA analysis for each of these two categories. But neither the audio tests nor the audio–video ones showed significantly different behaviours between musicians and non-musicians. This is directly attributable to many reasons. As a matter of fact, the tasks were not musical tasks but they involved everyday sounds, and everyone today has some experience with them even if non–musician. Furthermore, we did not ask to judge some musical features, but we only asked to evaluate generically their correspondence with a real sound. We will do a final consideration. Some of the Foley effects are very difficult to perform, because they need a deep experience in order to recreate them. Even if we trained in performing Foley sounds for two weeks, some Foley effects were not reproduced perfectly because we are no Foley artists. Definitely, this might have influenced the subjects in judging the life-likeness of each task.

4.2 Analysis of the audio data test

Each subject compiled a form questionnaire with different filling in the following fields: first name, gender, age, education and other questions such as “Do you usually listen to music?” and “Do you play any instrument? If so, how long have you been playing it?”. With these questions we were able to profile each participant as well as checking that they had a certain experience in listening critically to sound sequences. Mainly, we are going to analyze the principal ANOVA values, which are: the means, the F factor and the p-level. The F factor is simply the ratio of the two variance estimates. In the data that we will show next we had to omit all the results where the p- level was higher than the value 0.05 as they implied that the assumption that Foley sounds are different from real sounds was not true. In the audio test the action that had a p-level lower than 0.05 were:

- Walking on the summertime grass, which had a p-level of 0.001 an F factor of 12.71 and an average of 76.15 for the real sound while for the Foley sound is 53.05.
- Kiss, which had a p-level of 0.039 an F factor of 4.55 and an average of 54.50 for the Foley sound

Screen action	Preference
Walking on grass	Real 76.15
Kiss	Foley 54.50
Unkorking a bottle of wine	Foley 81.65
Passionate kisses	Foley 67.95

Table 2. Summary of the preferences for the actions with a lower p-level for the audio test.

and 37.30 for the real one.

- Uncorking a bottle of wine, which had a p-level of 0.018 an F factor of 6,08 and an average of 81.65 for the Foley sound while for the real sound is 65.95.

- Passionate kisses, which had a p-level of 0.020 an F factor of 5.85 and an average of 67.95 for the Foley sound and 47.25 for the real one.

The list above shows that only 40% of actions were completely distinguished. This is quite different from the results obtained with an average analysis (50%). Now we will draw up a list of the preferred sounds for each action with a low p-level for the audio test - referring to Table 1.

In our hypothesis the lack of important differences between Foley sounds and real sounds might be related to the fact that each sound expresses the action that it represents even if it has a different source. Therefore, we can assess if Foley sounds are equivalent to real sounds only with a numerical analysis of the signal, such as MFCC analysis or Onset detection.

4.3 Analysis of the Audio-Video data test

As the audio test, also the Audio-Video test lasted a week and employed twenty participants. Each of them compiled a questionnaire with the following fields: first name, gender, age, education and other questions such as "How often do you usually go to the movies?", "How many movies do you watch in a year?", "Do you play an instrument?". The purpose of the questionnaire was the same as the one of the audio test. Before analyzing the scores of the test we have to do a preliminary observation. First of all we discarded all the results from the listeners that evaluated the anchor sound as a very realistic sound. As a matter of fact, the anchor helped to discriminate with sufficient accuracy the participants that were not able to distinguish between different sound artifacts. Then we calculated the One-Way ANOVA for each action expected for the "Hand and Arm broken" because in this action twelve listeners out of twenty evaluated positively the anchor video. That might be due to the striking resemblance between the anchor sound, the Foley and real sound. Finally we kept the most significant data, which had a p-level lower than 0.05:

- Walking up the stairs, which had a p-level of 0.0196 an F factor of 13.29 and an average of 52.25 for the real sound while for the Foley sound is 28.50.

Screen action	Preference
Walking up the stairs	Real 52.25
Kiss	Foley 71.76
Passionate kisses	Foley 76.13

Table 3. Summary of the preferences for the actions with a lower p-level for the audio-video test.

- Kiss, which had a p-level of 0.0002 an F factor of 13.12 and an average of 71.76 for the Foley sound and 36.29 for the real one.

- Passionate kisses, which had a p-level of 0.00001 an F factor of 12.97 and an average of 76.33 for the Foley sound and 45.11 for the real one.

According to us, the lower score for the Foley sounds on the "Walking up the stairs" is due to the fact that the real sound has more features than the Foley sound, such as shoes noises or deeper reverberations, which allow to recognize it better. In this test only the 30% of actions were discriminated:

We can thus assert that there are no significant differences between a movie with Foley sounds and a movie with real sounds. As Michel Chion proposed in his book [4] the audio on a movie is only an added value to the pictures of the screen. We can demonstrate this hypothesis only through other tests that employ a higher number of subjects. However, even if our results highlight the fact that there are important differences between the audio and the Audio-Video test, it is not the main purpose of this paper to understand the psychological relationship between audio and video [5].

5. CONCLUSIONS

The main purpose of this paper was the direct comparison between Foley effects and real sounds, in order to understand if Foley sounds are always better than the real ones. What appears quite clearly observing and analyzing the findings is a similarity in judging preferences between real sounds and Foley sounds. As a concluding remark, we highlight the fact that the results of the tests demonstrate the participants partial discrimination ability between Foley effects and real sounds even though the sounds are remarkably different from each other. The final outcome of these experiments indicate a path of wider investigation on the world of Foley and everyday sounds. Therefore, future work will involve:

1. Further recording sessions of Foley sounds, so as to subdivide them in categories such as: impulsive sounds, continuous sounds, rhythmic sounds and so on.

2. Move from the realistic investigations of sounds to the evaluation of their expressivity.

3. A deep numerical analysis – with MFCC, centroid and many other methods – of the real and Foley sounds

in order to find out similarities or differences between them. In this way we aim at discovering which are the features that allow to recognize and better emphasize a sound.

4. Repeat both tests with more participants for each one and also with an equal number of female and male evaluators.

5. Understand how Foley effects exaggerate important acoustic features. These are the basis for being able to create a database of expressive sounds, such as audio caricatures, that will be used in different applications of sound design such as advertisement or soundtracks for movies.

Acknowledgments

We would like to thank Nicola Bernardini for his expert and enthusiastic support, Sergio Canazza for his useful advice for the practical tests and the *SaMPL* laboratory for having provided us with all the technical equipment needed for the recordings.

6. REFERENCES

[1] M. L. Heller and L. Wolf, "When sound effects are better than the real thing," *Journal of the Acoustical Society of America*, no. 111, p. 2339, 2002.

[2] A. V. Theme, *The Foley Grail: The Art of Performing Sound for Film, Games, and Animation*. Elsevier, 2009.

[3] G. Stoll and F. Kozamernik, "Ebu listening tests on internet audiocodecs," *Ebu Technical Review*, june 2000.

[4] M. Chion, *L'audio-vision. Son et image au cinma*. Paris: ditions Nathan, 1990.

[5] A. Kohlrausch and F. V. de Par, "Audio-visual interaction: From fundamental research in cognitive psychology to (possible) applications," *Human Vision and Electronic Imaging*, pp. 33–44, 1993.