REALITY (SOUND)BITES: AUDIO TRICKS FROM THE FILM AND TV STUDIO

Jay Rose, CAS

Digital Playroom
20 Marion St.
Brookline, MA 02446
jay@dplay.com

ABSTRACT

In the example-filled session that accompanies this paper, we'll listen to some of the ways sound designers fix—or sometimes, break—voices, music, and effects to help serve a director's vision. We'll start with how phoneme-level editing can change content, affect a dialect, or merge one voice with a completely different one. If time permits, we'll give some quick examples of how individual cookbook processes, such as equalization or delay, can be used in creative ways to change the nature of a sound. Finally, we'll examine how these processes are strung together in unusual ways, to simulate everything from an airplane interior to the sound of a classroom movie projector.

1. INTRODUCTION

Hollywood is full of phonies, losers, and fakes. Phony plastic props on a science fiction set don't sound like real anti-matter-based psychomegablasters. While boom mics are used to pick up natural-sounding dialog, they lose the tiny noises made by rustling clothing and footsteps. Movie fight scenes *have* to be faked—actors object to bruises and broken bones—so camera angles are cheated to make intentionally-missed punches look like they've connected [1].

By now, we've all seen *Making Of...* documentaries on TV, and realize that most movie sound effects have nothing to do with what was happening when the pictures were shot. Anti-matter devices are accompanied by sounds from synthesizers and samplers; and probably every moviegoer knows that a Star Wars light saber is actually the manipulated sound of striking a guy wire. Footsteps and clothing noises are recreated by *foley walkers* [2], specialists in the art of mimicking an on-screen actor's movements while sensitive microphones are pointed at their arms or feet. (If the actors are walking in snow, walkers may squeeze boxes of corn starch close to the mic; for walking in gravel, they might tamp their hands on bowls of cat food.) Foley walkers also fake fight noises, crashing heads of cabbage against a table for those haymaker punches and twisting bunches of celery for broken bones.

But that's the sexy stuff that provides good images for a documentary promoting a movie. There are more subtle manipulations in almost every film and TV soundtrack. When done right, you can't even hear that the sounds have been changed. Many of these techniques rely on principles that most ICAD members are familiar with... though you might not have thought about them in this context.

Figure 1. *Many movie sounds can be found at your local supermarket.*

2. EDITORIAL TRICKS

2.1. A Lost Art, Found Again

Television and film (and many home movies, these days) are cut in computers, letting the picture editor adjust *clips*—individual camera angles—on a frame-by-frame basis with the click of a mouse. Television and film require absolutely rock-steady frame rates, so picture editing software forces each edit to a one-frame temporal resolution. But a frame ($1/24^{th}$ second for film, approximately $1/30^{th}$ second for US television) is an enormously long time for spoken dialog. Some short sounds such as /t/ or /d/ may barely last 10ms (about $1/3^{rd}$ frame). Clean edits between one dramatic take and another, or from one thought to another in an interview, may be impossible in a frame-based world.

Back in analog days, film sound was recorded at first as an optical pattern on high-contrast black-and-white film (more about this in section 2.3), then later on magnetic stock; both had the same measurements and ran at the same speed as the picture negative [3]. 35mm movie film has four sets of sprocket holes for each frame of picture, so picture editors had to determine which of the four started a new frame; that was the only place they could physically edit. But even though the soundtrack had to be synchronized with the

picture, audio recordings don't have frame lines. Clever sound editors quickly learned they could cut on any arbitrary set of holes, and thus get 1/4-frame edit resolution. With this technique, an audio edit could be accurate to within 10.41ms. (Of course, non-sync audio on conventional magnetic tape can be cut at any arbitrary point. But this paper is about sound for *picture.*)

Figure 2. *35mm motion picture film has four sets of sprocket holes per frame. The accompanying soundtrack matches this format, but doesn't need to be cut on the frame line.*

Videotape made these quarter-frame edits impossible. Originally, the tape was physically spliced; sound and picture had to be cut together on a frame line. Later, the medium was edited by copying scenes from one tape to another under computer control. Under this scheme, audio and video could have different edit points, but edits were still restricted to frame boundaries. Even the elaborate CMX CASS [4] sound-editing computer system I used in the mid-1980s had that limitation: if I needed better than 1-frame resolution, I had to disable computer control and cut the tape with a razor blade.

Today's nonlinear video editing systems (NLEs) digitize the picture and sound, and store it on hard drives. The software nudges each video edit to the tiny slice of time between whole frames, to prevent playback instability. But while the linear PCM audio files used by the NLEs don't have frames and could theoretically be edited between any arbitrary set of samples, system designers have stuck to a model that makes one frame the smallest possible unit, even for sound editing.

Fortunately, audio software designers don't have this mindset. Modern digital audio workstations (DAWs) can have an edit resolution as fine as a single sample—about .02ms at the 48 kHz sample rate used for most film and video—and still maintain sync with a moving image. Many of the people editing dialog today come from either music or computerized video editing, and are uncomfortable thinking in terms smaller than musical beats or video frames. But we old-timers (and a new generation of hip sound editors who bother to study the art and its history) prefer to cut using a different measurement: phonemes.

2.2. The Glory of Phonemes

I have to assume that most ICAD members have studied phonetic transcription, or are at least familiar with its principles: any speech can be broken down into a small number of predictable sounds, based on how they're created by the throat and mouth, and these sounds then combine to form what we hear as syllables and—with occasional pauses—as words and sentences.

What phonetics teaches us to do with our ears and minds, can also be done in an audio workstation. Phonemes can be isolated and rearranged to fix a damaged speech recording, or even to change its meaning.

2.2.1. A quick example

I was demonstrating editing techniques a convention of the National Association of Broadcasters a few years ago. At that time, a particular sound clip was in the news: /aɪdɪd'nat hæv seks wɪθ'ðæt'wʊmæn/. (Recognize it?)

Radio stations were then parodying this soundbite by cutting out the third word. Well, *duh:* a brain-dead edit like that gives you an unnatural stress pattern, as well as language that's unusually formal for this context. As part of my demo, I sat at an Orban Audicy [5] DAW and in less than a minute changed the first few words to the much more natural-sounding /aɪhæd seks.../. I'll play both versions at the presentation.

While this trick relies on the obvious shifting of a few phonemes, what made me able to do it so quickly and smoothly was the presence of stop consonants (for this edit, /d/ and /t/) around some of the edit points Each has a tiny moment of silence which can be spotted quickly and can hide a transition between phonemes. Good dialog editors use a set of rules for editing different kinds of phonemes; I spend the better part of a chapter on them in my book *Audio Postproduction* (see section 6, Sources).

2.2.2. Presto, sex-changeo

Here's another phoneme trick. It's reasonable to assume we tell voices apart by the pitch and timbre of the underlying vocal buzz, and how it's filtered by head resonances, as well as content and context. But unvoiced consonants, in and of themselves, don't have any of these characteristics... which makes them powerful tools for dialog editing.

In my presentation, I'll play a clip of my wife and me:

Carla Have you read Jay's book?

Jay: Uh, actually I've written a couple of books.

Carla is a warm alto; I'm usually a baritone: you shouldn't have any trouble telling our voices apart. But if all you can hear is unvoiced consonants, our voices are identical. At the presentation I'll make one quick edit—putting my /ks/ in place of her /k/—and correct her statement without making the voice sound manipulated.

This is admittedly a "cooked" demonstration, relying on a recording made for the occasion, and also of the stop /k/ as an edit point. But the same technique is used frequently by dialog editors. In the presentation I'll play a real-world example from a commercial I cut for CBS Chicago last fall. It involves rock singer Janis Joplin, an "F-word", and my voice.

2.3. Back to that "Optical Pattern"

If you grew up with current technologies, it may surprise you that optical recording was used for motion picture soundtracks as long ago as the mid 1920s. But the pattern on the film had nothing to do with the pits on a CD-ROM or black-and-white spots of other optical media. These were *analog* recordings: a small metal vane would be linked to a voice-

coil, and moved in response to the changing voltage of the soundtrack. It constantly changed the area of a beam of light falling on film as the sound was recorded. When developed, the resulting pattern was an oscillogram, directly reflecting the audio waveform (figure 3). Skilled sound editors could recognize specific sounds on a track by eye, though they relied on a *sound reader* for the most accurate editing.

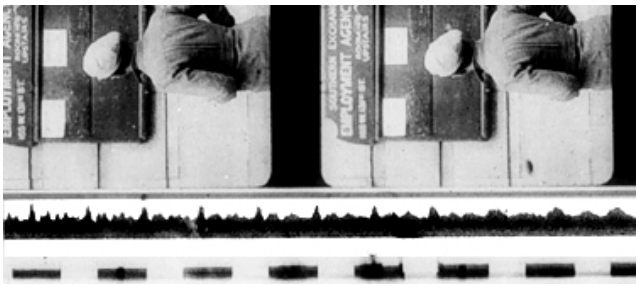

Figure 3. *An early variable-area optical soundtrack.*

Sound readers and theater projectors relied on the same mechanism to recover audio: a spot of bright light was focused on the moving soundtrack, shining through it and striking a photoelectric tube. The tube created a voltage that varied with the size of the clear part of the track, which was then amplified and sent to the speakers. (An earlier technology exposed the full width of the soundtrack at all times, but varied the transparency of the film according to the soundtrack. It was read the same way.)

There was no coding or decoding of data in a modern sense; changing voltage was turned into changing light, and then back again. The technique survives today: even digital theatrical soundtrack formats include a modern version of the variable area optical analog track as backup.

3. BASIC PROCESSING

When I started in this business [6], most recording studio consoles had volume knobs and on/off switches, and occasionally a way to add echo to selected microphones. That was it. There was no on-board equalization or elaborate routing. If you wanted equalization or other rudimentary effects, you had to patch through outboard equipment.

Two years ago I bought a small mixer for my studio: each of its 32 separate digital input channels has, in additional to motorized sliders for volume control and at least 16 different ways to route the signal, a 4-band parametric equalizer, a compressor / expander / noise gate, and a delay of up to half a second. That small mixer cost less than five thousand dollars. Top music studios invest hundreds of thousands in gigantic consoles, with even more on-board effects any myriad ways to use external DSP-based processors. That's the heart of the "studio magic" that turns eye candy into pop-music divas.

We use similar processing in film and TV sound; not to make an actor's voice more musical, but to give it more strength or reduce environmental noise. We also apply the processes in more radical ways to change the nature of sounds and create special effects.

3.1. Processing to Extremes

A couple of examples may prompt you to play with similar radical processing when designing auditory icons:

3.1.1. Dynamics control

The basic compressor (i.e., the device used to control dynamics; not the similarly-named codec for reducing data in an audio stream) smoothes out unevenness in a signal's envelope. It examines the incoming envelope and constantly changes its own gain based on user-set parameters. Compression is an essential part of telephony and speech recognition. Large amounts of the stuff—everything loud, all the time—were a hallmark of early Top 40s broadcasting. It helped AM radio signals fight the limitations of shirt-pocket and car radios. Smaller amounts are used on almost all television dialog, for a similar reason.

Sometimes the signal is split into separate bands and each is compressed simultaneously; this *multiband compression* is one of the secrets behind those loud, dense movie trailer soundtracks. When taken to extremes, it provides the distorted timbres and constant high-frequency energy of much pop music.

But even simple compression can be used creatively to change the nature of a sound effect. This is done by manipulating the compressor's time constants: the *attack* determines how long it takes the circuit to react after the initial attack of a loud sound; the *decay* controls how long it takes the gain to return to normal after the sound gets soft again. These adjustments are rarely touched in speech processing, but can have profound effects on music or sound effects.

Figure 4A shows the envelope of an unprocessed sound, a .357 Magnum shot at medium range. 4B is the same sound with a slow attack and decay; only the initial sound gets through, and the result sounds more like a pile-driver's clank than a bang. 4C has a very fast attack and decay: the hit is suppressed but the natural reverb is emphasized, making it sound more like a big explosion. Most people hearing this clip swear it's from three different sound sources.

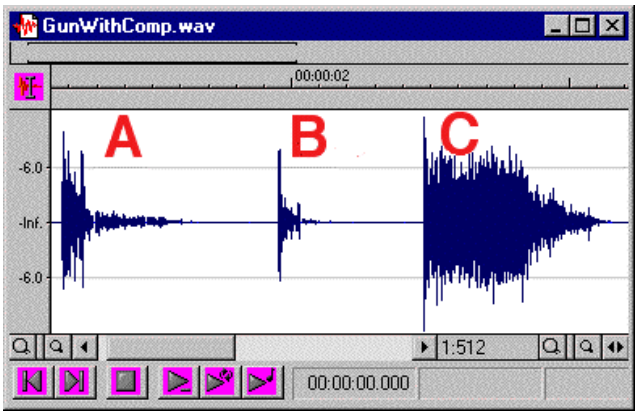

Figure 4. *Three versions of a .357 Magnum shot; the only differences are how the time constants of a compressor have been changed.*

3.1.2. Equalization as Effect

We're all familiar with the "telephone filter" used as a special effect in radio commercials and some pop music. This is a fairly straightforward process using a pair of fourth-order or higher cutoff filters around 350 Hz and 4 kHz, and is in every sound designer's arsenal.

A more novel use of equalization depends on the tendency of a peaking filter with a very high Q (narrow bandwidth) and moderate gain to go into oscillation when triggered by a signal at its design frequency. This is the electronic equivalent of the feedback howl heard when a public address system interacts badly with a room's acoustics.

One of my ongoing projects is designing software for a multi-DSP processor used to produce special effects in broadcast and film [7], and I decided to take advantage of this resonance effect for one category of sounds. Speech input is passed through multiple parallel peaking equalizers, set in the vocal formant region. The Q and gain of each equalizer is kept just under the threshold of oscillation; their frequencies are adjustable to provide musical chords in various keys.

The resulting effect can be described as a "talking harp", but without the obvious gating that a vocoded recording of an actual harp would produce. I'll play an example at the presentation; maybe the attendees will think of a better name.

4. COMBINING PROCESSES

Approximately 125 separate algorithms of mine will have been released in this software project by ICAD 2003. Almost all of them rely on basic DSP processes—equalizers, delays, oscillators, and various kinds of modulation—combined in unique ways. Designing them starts with a logical analysis of the sound I'm trying to create.

As a practical example, consider the frequent problem of trying to turn modern video into old-fashioned film, anything from a torn, dirty classroom film to a resurrected Hollywood classic. On the video side, the usual formula is to reduce color saturation (or shift the colors), add some scratches and dirt, and simulate some motion artifacts that would normally be caused by converting 24 fps film into 29.97 fps video. This combination is so common, it's a preset in many NLEs. Of course it's a parody of film, emphasizing its limitations to broaden the gag.

For an audio equivalent, I decided to design a process that would parody a film's soundtrack. Analog optical recording, described in section 2.3 and used in some way on virtually every film ever made, has plenty of limitations to play with.

The following paragraphs show what's involved in designing a multi-process effect. The bulleted paragraphs reflect the analysis; the non-bulleted ones are the practical steps that result.

So: what makes the characteristic sound of, say, a classroom projector?

- The projector has to hold the film perfectly still, 24 times a second, so individual frames can be projected. But the film has to be in constant, smooth motion a few inches away because the sound head needs a continuous signal. These two motions are isolated by a shock-absorbing loop of film between the lens and the sound head. But the isolation was never perfect, and some of that 24 frame per second flutter would always sneak in.
- In badly-maintained projectors, the mass of the film reels influences the speed. If the mass isn't perfectly balanced—as can happen when the reels are warped — the speed keeps changing as the reel moves.

To get this sound, frequency-modulate the input with a 24 Hz sinewave (that'll be the flutter) combined with about a .2 Hz sinewave (wow from a 1200' reel of 16mm film). The easiest way is to mix two oscillators and let the result modulate the length of a delay. As the length changes, it forces the pitch up and down. Using a delay of about 40 ms also knocks the track one film frame out of sync, which may add to the humor of the effect.

- Classroom projectors also had a characteristic hum. The lamp used to pour light through the optical soundtrack needed a pure DC voltage; any periodic variation in the voltage would be amplified with the track. It was hard to generate this voltage in a badly-maintained projector, and some of the power line's AC would sneak in.

Add a 120 Hz sawtooth, at low levels, for the hum.

- The tube that read the soundtrack had its own random noise.

Add some white noise.

- Optical sound did not have a wide frequency range. Both high and low frequencies were lost in the process. Bad projector maintenance would destroy the highs even more.

Bandpass everything between 250 Hz – 7.5 kHz. The lower filter should be fairly sharp. The upper one depends on what you're trying to simulate. Well-maintained movie theaters running 35mm tracks could started to fall off around 12 kHz, but a sharp 7.5 kHz filter will make sure the effect is immediately recognizable to an audience. Classroom projectors ran at a slower speed, were not high fidelity to start with, and lost treble when they got out of alignment... use a filter that starts a gentle roll off around 2.5 kHz. Before filtering, throw in a little distortion by hard-limiting the signal.

- It isn't really part of the track, but the clicking noise from a projector's gate is part of the cliché.

Use a 12 Hz square wave, high-pass filtered very sharply at 5.5 kHz. Modulate the frequency slightly with the .2 Hz sinewave used for wow, and it's surprisingly realistic.

- Classroom projector motors had to fight a lot of mass and always got off to a slow start.

Use a pitch bender to ramp both speed and pitch up from zero to normal over about three seconds when the clip starts, and back down to zero as it's ending.

- Giant movie palaces had a sound of their own. A 75' deep auditorium had a distinct echo from the rear wall, about 2.5 frames after dialog left the speakers. Every theater was slightly different, and the echo varied depending on where you were sitting

Use a reverb with a short, diffuse characteristic. Set the time before the first reflection to about 100 ms, and bunch the other early reflections together.

In the commercial implementation of this algorithm, I added an optional "splice". Users can press a button to create

a loud click and jump forward in the track (it works by adding a slowly-grown delay to the realtime input, then canceling the delay after the click).

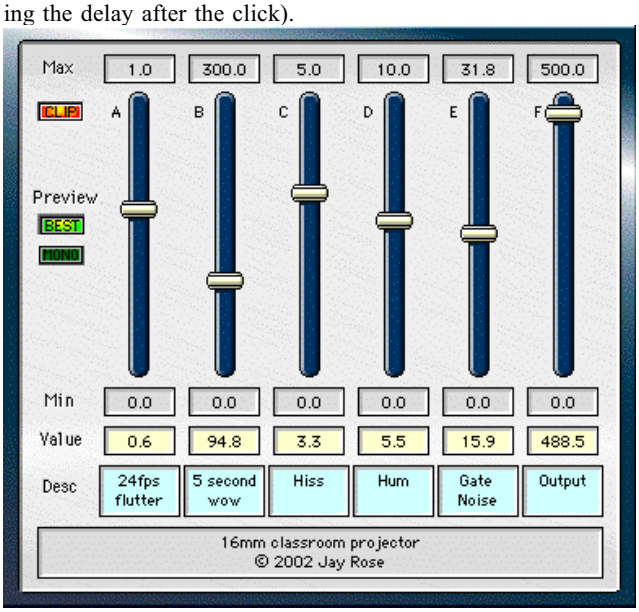

Figure 5: Most of the functions of the projector simulation are implemented in this free preset file.

I also released a simplified freeware version as a tutorial (Figure 5), with all of the above features except for the pitch bend, reverb, and splice. It's a preset for SFX Machine, a $25 swiss-army audio processor that every Mac-equipped sound designer should own. You can download my preset and get information about its host software at www.dplay.com/dv, entry for June '02.

5. SOURCES

The examples and sounds in this presentation are taken from two books of mine. Additional details on both, along with tables of contents, downloadable sample text, critical and reader comments, and links to the best discount prices I've found on the web are at `www.dplay.com/book`.

5.1. Producing Great Sound for Digital Video, 2nd Edition

(CMP Books, DV Expert Series, 2002) 420 pages with audio CD. Covers the entire soundtrack process, with particular emphasis on planning for sound and gathering dialog at the shoot. Used in many film classes; a free Teacher's Guide is available through the website.

5.2. *Audio Postproduction for Digital Video*

(CMP Books, DV Expert Series, 2002) 430 pages with 1-hour audio CD, $44.95. Goes into much deeper detail of what happens to a soundtrack after the shoot, with explanations of how the most important processors actually handle the sound as well as how to apply them, and cookbook recipes for common situations.

6. ACKNOWLEDGEMENTS

Thanks to my son Dan Rose, Chief Operator of radio station WBUR, and audio software and loudspeaker engineer Richard Pierce, for helping me sort through some of the concepts I'm explaining. Thanks especially to Professor Barbara Shinn-Cunningham of Boston University's Departments of Cognitive and Neural Systems and Biomedical Engineering for critiquing early versions of this paper, and for helping me understand the academic's perspective.

7. REFERENCE AND NOTES

[1] Rose, Jay "Walk This Way: The Art of Foley". *Digital Video Magazine* August 2002, p 82, for more on this topic.

[2] The name honors Jack Foley, a Hollywood second-unit director and editor. He didn't invent the technique, but in the late 1940s he popularized it among producers.

[3] Magnetic recording stock that was 35mm wide used a lot of oxide solution. To save money, studios often used the same transparent base as their movie film but with a narrow magnetic stripe instead of photographic emulsion.

[4] It controlled a videotape deck, 24-track 2"-tape audio recorder, and two 1/4" audio source decks from an ASCII keyboard, using SMPTE timecode tracks on each tape for synchronization and precision.

[5] Which I still use. Its design was spearheaded by ICAD 2003 treasurer Barry Blesser.

[6] It was 1969, and my graduate assistantship was programming and producing for a university's public radio station.

[7] Eventide DSP4000B+ and Orville series of processors; http://www.eventide.com/profaud/dsp4000b.htm. My assignment is to create special effects for broadcast and film (and it's gratifying to hear them being used in network programs and in movies), but the software has also been adopted by music remixers. Go figure.