SocialFX: Studying a Crowdsourced Folksonomy of Audio Effects Terms

Taylor Zheng
Northwestern University
EECS Department
tz0531@gmail.com

Prem Seetharaman
Northwestern University
EECS Department
prem@u.northwestern.edu

Bryan Pardo
Northwestern University
EECS Department
pardo@northwestern.edu

ABSTRACT

We present the analysis of crowdsourced studies into how a population of Amazon Mechanical Turk Workers describe three commonly used audio effects: equalization, reverberation, and dynamic range compression. We find three categories of words used to describe audio: ones that are generally used across effects, ones that tend towards a single effect, and ones that are exclusive to a single effect. We present select examples from these categories. We visualize and present an analysis of the shared descriptor space between audio effects. Data on the strength of association between words and effects is made available online for a set of 4297 words drawn from 1233 unique users for three effects (equalization, reverberation, compression). This dataset is an important step towards implementing of an end-to-end language-based audio production system, in which a user describes a creative goal, as they would to a professional audio engineer, and the system picks which audio effect to apply, as well as the setting of the audio effect.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human factors; H.5.1 [**Multimedia Information Systems**]: Audio input/output; H.5.2 [**User Interfaces**]: User-centered design; H.5.5 [**Sound and Music Computing**]: Signal analysis, synthesis, and processing

Keywords

Interfaces; audio engineering; effects processing; signal processing; reverberation; equalization; compression; vocabulary; crowdsourcing

1. INTRODUCTION

Audio production is a critical part of the professional production of many forms of media. Audio production tools, such as reverberation, equalization, and compression, are used to process audio after it is recorded, transforming these raw recordings into polished final products. When communicating audio production goals in these settings, content creators often use language as the primary communication medium. Meaningful language is needed when communicating these goals, since the language used in this context has connotations that are particular to audio production tools.

People with little or no training on audio production tools often describe their creative audio goals with vocabulary that has no obvious path to realization using given audio production tools. Many such potential users of audio production tools (e.g. acoustic musicians, podcast creators) have sonic ideas that they cannot express in technical terms. They may not even be able to say which audio effect tool is used to achieve their goals. As a result, they have difficulty using such tools and interactions between audio production professionals and these content creators can be a frustrating experience, hampering the creative process.

The following quote from Jon Burton, of *Sound on Sound*, illustrates the communication problem audio engineers face:

"...how can you best describe a sound when you have no technical vocabulary to do so? It's a situation all engineers have been in, where a musician is frustratedly trying to explain to you the sound he or she is after, but lacking your ability to describe it in terms that relate to technology, can only abstract. I have been asked to make things more 'pinky blue', 'Castrol GTX-y' and 'buttery'." [1]

In this work, we study a vocabulary that a population of non-experts in audio engineering produced to describe audio effects produced by three of the most widely used effects tools: equalization (EQ), reverberation, and dynamic range compression (compression). Equalization adjusts the gain of individual frequencies in a recording and can be used to make things sound brighter or warmer. Reverberation adjusts the spatial quality of an audio recording by adding echo effects to the audio and can be used to make things sound like they were recorded in a cave, or a church, or a stairwell, etc. Compression reduces the dynamic range of an audio recording by reducing the amplitude of parts of audio above a specified decibel value and can be used to increase the sustain of instruments, reduce sibilant and plosive frames of a vocal recording, and prevent clipping when multiple tracks are mixed together. The EQ and reverberation datasets were described and presented in [2] and [3]. This work adds another dataset consisting of a vocabulary for compression, describes a general framework for obtaining vocabularies for arbitrary audio effects, and makes a dataset available to the public for equalization, reverberation, and compression.

In this work, we consider the following questions:

MM '16, October 15-19, 2016, Amsterdam, Netherlands
© 2016 ACM. ISBN 978-1-4503-3603-1/16/10. . . $15.00
DOI: http://dx.doi.org/10.1145/2964284.2967207

1. How can we discover words used by laymen to describe arbitrary audio effects?
2. What words are associated with which audio effects?
3. What words are associated with audio effects in general, and can be achieved effectively using multiple audio effects?

We extend existing work ([2], [3]) into SocialFX, a crowdsourcing solution for discovering words used by a target population for describing an arbitrary audio effect. Importantly, this data collection doesn't merely collect words, it maps words onto concrete manipulations performed by EQ, reverberation and compression tools to make an actionable vocabulary that can be used to create effects tools to manipulate sounds in terms of these words. We look at the data collected using this approach to examine how words are used across multiple audio effects, offering the first insights into the shared descriptor space of audio effects.

2. RELATED WORK

There are several existing works for learning descriptors for audio. One common approach is to use text co-occurence, lexical similarity, and dictionary definitions (e.g. Wordnet [4]. These approaches are not sufficient, as we wish to examine the mappings between words and measurable sound features and controls for audio effect tools.

Psychologists have studied the mappings between descriptive terms and measurable signal characteristics for sound. Some terms, such as those for pitch (high, low) or loudness (loud, soft), have well defined mappings onto sounds [5], [6]. Others, such as “underwater”, or “muffled”, have no obvious connection onto audio tools. There have been numerous attempts since the 1950s that hope to find universal sound descriptors that relate to a set of canonical perceptual dimensions ([7], [6], [8], [9]). In recent years, researchers from many different backgrounds, such as recording engineering [10] [11], music composition [12], and computer science [13] have tried to find a universal set of descriptive terms for sound.

In [14], audio features are extracted from recordings from onomatopoeia and mapped into a perceptual space, where distance between terms is correlated with perceptual distance. This work focuses on onomatopoeia, rather than the broader range of all possible audio effects, and on a small population of four lab members, rather than the larger lay population.

In [15], [16], a reverberator is developed that can be controlled entirely through perceptual characteristics of the signal, rather than in terms of low-level audio signal processing. However, these works are limited to just a few words selected by the researcher, and is limited to reverberation. Our work finds many more words as elicited from a population of laymen, and works for arbitrary audio effects.

In [2] and [3], two distinct approaches to collecting effects vocabulary data were followed, with both approaches utilizing Amazon Mechanical Turk to crowdsource effect descriptor data. SocialEQ first asked users to provide a descriptor word, then to select one of three audio samples. The selected audio would have an effect applied to it (in this case, EQ), and users were asked to rate how well the resulting audio fit the descriptor they supplied. After 40 ratings, the system would have enough data to construct an effect with parameters that fit the supplied descriptor, resulting in a mapping of an effect's parameter space to a descriptor space over the course of many sessions.

In contrast, SocialReverb asked users to listen to an audio clip randomly chosen from a group of three clips, first with no effect applied and then with an effect applied, with parameters randomly chosen from a pool of 256 parameter configurations as specified in [2]. Users were then asked to describe the resulting effect, first with as many words as they freely desired, then with descriptors they agreed with, chosen from a pool of previously contributed words. Users then rated how strongly the applied effect affected the audio clip on a Likert scale. Much like SocialEQ, the resulting data maps the parameter space of an effect to a descriptor space over the course of many sessions. For our work, we chose to follow the approach used in SocialReverb, replacing reverberation with compression.

Taking into account exclusion criteria listed in [3], the data for Social-EQ was collected in 731 sessions from 481 individuals, resulting in a pool of 324 unique descriptors for equalization. Similarly, taking into account exclusion criteria listed in [2], the data for SocialReverb was collected from 513 individuals describing 256 unique instances of reverberation parameter configurations, resulting in 2861 unique descriptors for reverberation.

3. SOCIALFX

We build directly on the work in [2] and [3], extending it to SocialFX, a system for collecting descriptors for arbitrary audio effects from a population of laymen. In this work, we collect data on a new audio effect, compression. Then, we combine our compression vocabulary with the vocabularies previously collected for reverberation and EQ so as to analyze the relationships between descriptor spaces for these three different audio effects.

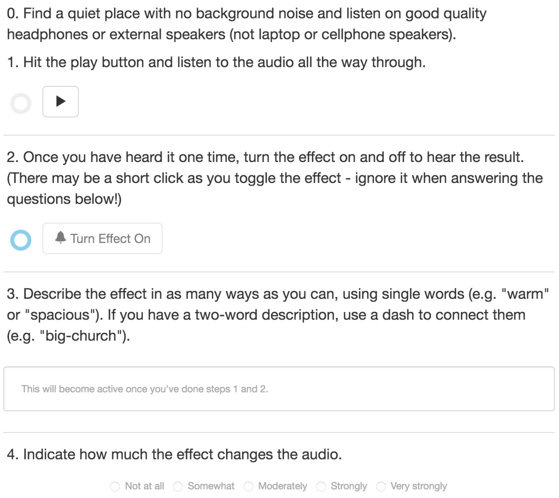
We used Amazon Mechanical Turk and the interface in Figures 1 and 2 to crowdsource data on how people describe compression. Taking into consideration exclusion criteria similar to that used in [2], our data was collected from 239 individuals describing 256 unique instances of compression parameter configurations, resulting in 1,112 unique descriptors.

When analyzing the shared descriptor space across EQ, reverberation, and compression, we were interested in learning how strongly a descriptor is associated with each effect. Both audio effects experts (recording engineers) and non-experts (acoustic musicians, podcasters, videographers, etc) reach a shared understanding of what effect one is talking about when using a specific descriptor, reducing misunderstandings in the creative process.

To determine the particularity of a descriptor, we first calculate the frequency of appearance of a descriptor within an effect by dividing the number of occurrences of that descriptor within an effect by the total number of descriptor instances in the data set of that effect. Then we divide the descriptor space according to whether a descriptor is shared among all three effects or not. The descriptors in common with all three effects are further divided depending on how frequently they occurred for each effect; if a descriptor appeared with high frequency for reverberation but with low frequency for EQ and compression, we can conclude that the descriptor leans toward reverberation, while if a descriptor appeared with roughly equal frequency among all three effects, it is a more general descriptor.

Word category	EQ	Reverberation	Compression
General words	warm, loud, soft, happy, cool, clear, muffled, sharp, bright, calm, tinny		
Tending words	cold, happy, soothing, harsh, heavy, beautiful, mellow	distant, deep, hollow, large, good, grand, spacey	quiet, full, sharp, crisp, energetic, sutble, clean, fuzzy
Specific words	chunky, wistful, punchy, mischievous, aggravating	haunting, organ, big-hall, church-like, concert, cavernous, cathedral, gloomy	volume, sharpened, feel-good, rising, peppy, easy-going, earthy, clarified, snappy

Table 1: Descriptors and which audio effect they are related to. General words are used to describe audio effects produced by any of the three effects tools. Tending words are ones which were shown predominantly for a single audio effect, but appear in other audio effect vocabularies with low frequency. Specific words are ones that are used for a single audio effect and no others. The words shown above were found via inspection of the shared descriptor space between the three audio effects. The general words can be seen in Figure 3.



0. Find a quiet place with no background noise and listen on good quality headphones or external speakers (not laptop or cellphone speakers).

1. Hit the play button and listen to the audio all the way through.

▶

2. Once you have heard it one time, turn the effect on and off to hear the result. (There may be a short click as you toggle the effect - ignore it when answering the questions below!)

Turn Effect On

3. Describe the effect in as many ways as you can, using single words (e.g. "warm" or "spacious"). If you have a two-word description, use a dash to connect them (e.g. "big-church").

This will become active once you've done steps 1 and 2.

4. Indicate how much the effect changes the audio.

○ Not at all ○ Somewhat ○ Moderately ○ Strongly ○ Very strongly

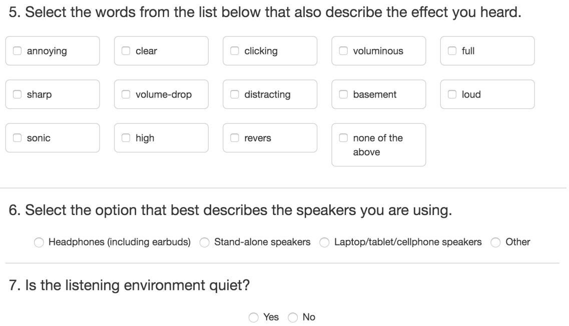
Figure 1: Part one of SocialFX: Participants are asked to listen to a dry recording, then a recording with an audio effect applied, and then describe it in their own words.

We end up with three general categories of descriptors: ones that are specific to an effect, ones whose usage leans toward a particular effect, and ones that are general across all three effects. Examples of these are shown in Table 1.

4. DATASET ANALYSIS

Figure 3 visualizes the shared descriptor space across all three effects, with each axis representing the frequency of occurrence of each shared descriptor within the data set for each audio effect. In the shared descriptor space, we see certain words such as *warm* or *loud* are used broadly across compression, equalization, and reverberation, while other words, such as *soothing* or *full* tend towards one audio effect.

Within the shared data set, there are generic words such as *sound* and *normal* that have no strong connotations or associations with a particular effect. On the other hand, words such as *dark*, *bassy*, *tinny*, *bright*, and *warm* are all strongly associated with EQ. Their appearance as descriptors in both reverberation and compression can be explained by the fact that these two effects can alter the equalization of audio; in some cases, reverberation and compression can reduce the high frequency content of audio, leading to descriptors such as *warm* and *dark*.



5. Select the words from the list below that also describe the effect you heard.

□ annoying □ clear □ clicking □ voluminous □ full
□ sharp □ volume-drop □ distracting □ basement □ loud
□ sonic □ high □ revers □ none of the above

6. Select the option that best describes the speakers you are using.

○ Headphones (including earbuds) ○ Stand-alone speakers ○ Laptop/tablet/cellphone speakers ○ Other

7. Is the listening environment quiet?

○ Yes ○ No

Figure 2: Part two of SocialFX: After completing part one of SocialFX, participants are asked to look at a set of words that other people used to describe the same audio effect, and check off which ones they agree describe the effect.

Words that are usually associated with reverberation also appear in the list of common descriptors, such as *distant* and *spacey*. This can be explained in the case of EQ by the fact that reducing mid-range frequencies relative to treble and bass frequencies can create a greater perceived sense of distance from an audio source.

Smooth and *even* are words usually associated with compression that were used to describe EQ and reverberation as well. EQ and reverberation can potentially be used to reduce sibilants and transients in audio tracks, which can be perceived as *smooth* or *even*. Words like *quiet*, *soft*, and *loud* also all deal with volume levels, but can be achieved via reverberation by reducing the amount of direct sound or via equalization by damping prominent frequencies.

The list of shared descriptors also has bridge words, which are words that have different meanings in different contexts. For example, *hollow* in the context of EQ usually refers to a lack of mid-range frequencies, while in the context of reverberation, it can refer to the feeling of space generated by reverberation. *Crisp*, for EQ, refers to an abundance of upper treble frequencies, but for compression, it can refer to the preservation of transients under subtle compression settings. We find that the vocabularies of the three audio effects are often intertwined.

5. DATA SET

To facilitate the creation of word-based interfaces that use non-expert vocabulary to control audio production tools,

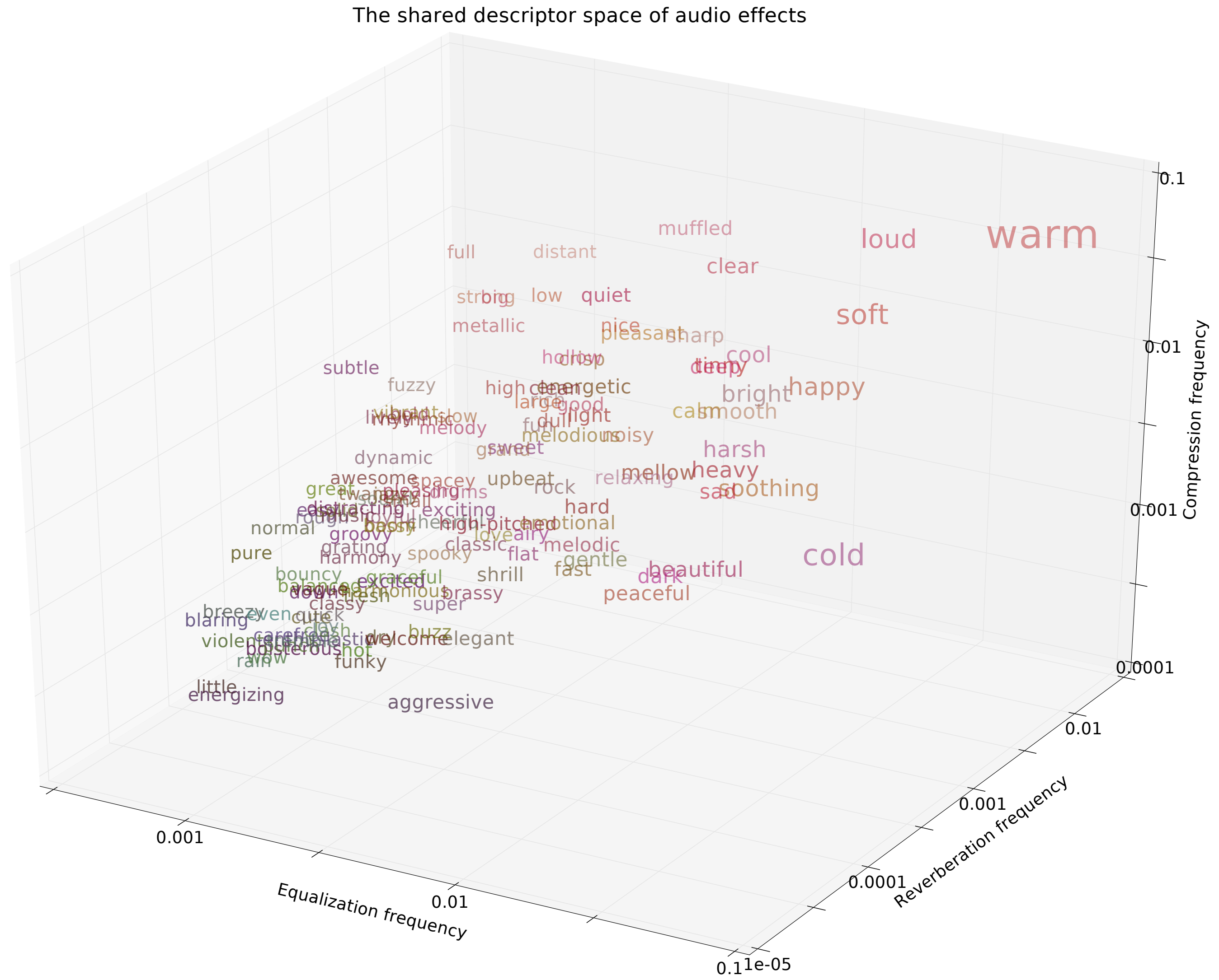


Figure 3: Shared descriptor space arranged in terms of frequency of occurrence in each effect data set. Towards the top right indicates high frequency across all audio effects (e.g. *warm*). The size of the word correlates with how often it was used across all three datasets. Words that tend towards an effect can be visualized along each axis. As words tend along the reverberation frequency axis, they become more transparent and become more red, to make the 3D effect easier to see.

we have created a data set which we will make available at http://bit.ly/1WmTP6v. The data set includes relative word frequency of 4297 words drawn from 1233 unique users across three effects (EQ, reverb, compression), as well as the associated effects settings. We also plan to develop a Javascript library for the development of language-based audio production interfaces.

6. CONCLUSION

In this work, we have presented SocialFX, a crowdsourcing mechanism for discovering vocabulary related to audio effects. We have presented an analysis of three datasets, each collected for different audio effects - equalization, reverberation, and compression. We have found that there are three categories of words used to describe audio: ones that are generally used across effects, ones that tend towards a single effect, and ones that are exclusive to a single effect. We have shown examples of these three categories. Finally, we have visualized and presented an analysis of the shared descriptor space between audio effects. Our analysis of these descriptor spaces shows a way forward to alleviate communication difficulties in audio production environments that are caused by the use of language.

This analysis is a first step toward an end-to-end language-based audio production system, in which a user describes a creative goal, as they would to an audio engineer, and the system picks which audio effect to apply, in addition to adjusting that effect's parameters to achieve the user's goal.

7. ACKNOWLEDGMENTS

We would like to thank NSF Grants 1116384 and 1420971 for funding this work. Thanks to Alison Wahl for providing source audio for SocialFX.

References

[1] Jon Burton. *Ear Machine iQ: Intelligent Equaliser Plug-in.* June 2011. URL: http://www.soundonsound.com/sos/jun11/articles/em-iq.htm.

[2] Prem Seetharaman and Bryan Pardo. "Reverbalize: a crowdsourced reverberation controller". In: *ACM Multimedia, Technical Demo* (2014).

[3] Mark Cartwright and Bryan Pardo. "Social-eq: Crowdsourcing an equalization descriptor map". In: *14th International Society for Music Information Retrieval.* 2013.

[4] George A. Miller. *WordNet: a lexical database for English.* 1995. DOI: 10.1145/219717.219748.

[5] H. Helmholtz and A. Ellis. *On the sensations of tone as a physiological basis for the theory of music.* Dover, New York, 2nd english edition, 1954.

[6] S. McAdams et al. *Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes.* Psychological Research, 58(3):177-192, 1995.

[7] J. Grey. *Multidimensional perceptual scaling of musical timbres.* The Journal of the ASA, 61(5):1270-1277, 1977.

[8] L. Solomon. *Search for physical correlates to psychological dimensions of sounds.* The Journal of the ASA, 31(4):492-497, 1959.

[9] A Zacharakis, K Pastiadis, and G Papadelis. "An Investigation Of Musical Timbre: Uncovering Salient Semantic Descriptors And Perceptual Dimensions". In: *12th International Society for Music Information Retrieval Conference.* 2011.

[10] D. Huber and R. Runstein. *Modern recording techniques.* Focal Press/Elsevier, Amsterdam ; Boston, 7th edition, 2010.

[11] Ryan Stables et al. "SAFE: A system for the extraction and retrieval of semantic audio descriptors". In: *15th International Society on Music Information Retrieval* (2014).

[12] D. Smalley. *Spectromorphology: explaining sound-shapes.* Organised Sound, 2(02):107-126, 1997.

[13] M. Sarkar, B. Vercoe, and Y. Yang. "Words that describe timbre: a study of auditory perception through lan- guage". In: *Proc. of Language and Music as Cognitive Systems Conference.* 2007.

[14] S Sundaram and S Narayanan. "Analysis of audio clustering using word descriptions". In: *ICASSP: Acoustics, Speech and Signal Processing* (2007).

[15] Zafar Rafii and Bryan Pardo. "Learning to Control a Reverberator using Subjective Perceptual Descriptors". In: *10th International Society on Music Information Retrieval* (2009).

[16] Zafar Rafii and Bryan Pardo. "A Digital Reverberator Controlled through Measures of the Reverberation". In: *Northwestern Electrical Engineering and Computer Science Department* (2009).