SOCIAL-EQ: CROWDSOURCING AN EQUALIZATION DESCRIPTOR MAP

Mark Cartwright
Northwestern University
EECS Department
mcartwright@u.northwestern.edu

Bryan Pardo
Northwestern University
EECS Department
pardo@northwestern.edu

ABSTRACT

We seek to simplify audio production interfaces (such as those for equalization) by letting users communicate their audio production objectives with descriptive language (e.g. "Make the violin sound 'warmer.'"). To achieve this goal, a system must be able to tell whether the stated goal is appropriate for the selected tool (e.g. making the violin "warmer" with a panning tool does not make sense). If the goal is appropriate for the tool, it must know what actions need to be taken. Further, the tool should not impose a vocabulary on users, but rather understand the vocabulary users prefer. In this work, we describe SocialEQ, a web-based project for learning a vocabulary of actionable audio equalization descriptors. Since deployment, SocialEQ has learned 324 distinct words in 731 learning sessions. Data on these terms is made available for download. We examine terms users have provided, exploring which ones map well to equalization, which ones have broadly-agreed upon meaning, which term have meanings specific small groups, and which terms are synonymous.

1. INTRODUCTION

Much of the work of audio production involves finding the mapping between the terms in which a musician describes an acoustic concept (e.g. "Make the violin sound 'warmer.'") and the tools available to manipulate sound (e.g. the controls of a parametric equalizer). Often, mappings are non-obvious and require significant work to find.

We seek to simplify audio production interfaces (such as those for equalization) by letting users communicate their audio production objectives with descriptive language ("Make the violin sound 'warmer.'"). To achieve this goal, the system must be able to tell whether the stated goal is achievable for the selected tool (e.g. making the violin "warmer" with a panning tool does not make sense). It must also know what actions need to be taken, given the correct tool ("Use the parametric equalizer to boost the 2-4 kHz and the 200-500 Hz bands by 4 dB"). Further, the tool should be aware of possible variations in the mapping between word and audio among users (Bob's "warm" $\neq$ Sarah's "warm"), and the tool should be aware of which words are synonymous.

In this work, we describe SocialEQ, a project to crowdsource a vocabulary of audio descriptors, grounded in perceptual data that can be mapped onto concrete actions by an equalizer (EQ) to effect meaningful change to audio files. SocialEQ has, to date, been taught 324 distinct words in 731 learning sessions. Through our analysis of the data collected, we address the following questions:

1. What audio descriptors are actionable by an audio equalizer?
2. How widely-agreed-upon is the meaning of an EQ descriptor?
3. What EQ descriptors are true audio synonyms?

2. BACKGROUND AND RELATED WORK

There is currently no universal dictionary of audio terminology that is defined both in terms of subjective experiential qualities and measurable properties of a sound. Tools built from text co-occurrence, lexical similarity and dictionary definitions (e.g. WordNet [13]) are fundamentally different in their underpinnings and do not address the issue of how words map to measurable sound features.

There are some terms relating to pitch (high, low, up, down) and loudness (soft, loud) that have relatively well-understood [6, 18] mappings onto measurable sound characteristics. Some terms of art used by recording engineers [8] describe effects produced by recording and production equipment, and are relatively easy to map onto measurable properties. These include "compressed" (i.e., a compressor has been applied to reduce the dynamic range of the audio) and "clipped" (i.e., the audio is distorted in a way characteristic of an overloaded digital recorder). These terms are not, however, widely understood by either musicians or the general public [21].

Numerous studies have been performed over the last fifty years in the hopes of finding universal sound descriptors that map onto a set of canonical perceptual dimensions [4, 11, 20, 22]. Also, in the last decade or so, many researchers coming from backgrounds such as recording engineering [8], music composition [19] and computer science [16] have studied the space of terminology, seeking a

universal set of English terms for timbre. Researchers in acoustic signal processing and perception continue to seek English taxonomies for timbre descriptors. Some seek general taxonomies of sound descriptors [16]. Others are focused on timbre adjectives for particular instruments [2,9].

Such studies typically start by determining a set of natural descriptors by performing a survey. These descriptors are then used in a second study, where participants evaluate sounds in terms of the descriptors. These are then mapped onto machine-measurable parameters, such as spectral tilt, sound pressure level, or harmonicity. Commonalities in descriptor mappings are found between participants in the studies and then some small set of descriptive terms are proposed as relatively universal.

In audio engineering, there has been work directly mapping equalizer parameters to commonly used descriptive words using a fixed mapping [12, 14]. A problem with these approaches is that mappings are very time-consuming to develop, the former requires a painstaking training process for each user, and the latter is brittle with respect to the variability that exists across users.

In contrast, [15] establishes the effectiveness of an approach to automatically learning mappings of audio adjectives onto actionable controllers in a small study using 4 researcher-selected descriptors and 19 participants. This study was, however, quite small in the number of participants and the number of adjectives learned.

Despite the efforts of all these groups, a universal map of descriptive terms for audio remains an unachieved goal. We believe this is because the terms used range from widely-agreed-upon (e.g. *loud*) to ones that have agreement within groups but not between groups (e.g. *warm*) to idiosyncratic terms meaningful solely to individuals (e.g. *splanky*).

In this work, rather than seek a set of universal underlying dimensions of timbre or the universal meaning of a descriptive terms in all auditory contexts, we seek an understanding of descriptors in a specific context: audio equalization.

3. THE SOCIALEQ TASK

`SocialEQ.org` is a web-based application that learns an audio equalization curve associated with a user-provided audio descriptor. To deploy the software to a large audience for our data collection, we have implemented it as a web application using Adobe Flex and Drexel University's Audio Processing Library for Flash (ALF) [17].

After agreeing to participate, we ask participants to "enter a descriptive term in the language in which you are most comfortable describing sound (e.g. 'warm' for English, 'claro' in Spanish, or 'grave' in Italian), pick a sound file which we will modify to achieve the descriptive term, then click on 'Begin'."

Participants are then given the choice of three different source audio files to manipulate: "Electric Guitar", "Piano", and "Drums". All files were sampled at 44.1 kHz and 16 bits. The files all had constant sound (i.e. no breaks in the audio) and were presented in loops that were 8, 10, and 14 seconds long, respectively.

Figure 1. SocialEQ ratings screen. The user rates how well each audio example corresponds to the descriptive term they selected at the prior stage. Users rate an example by dragging the associated circle to the desired horizontal position. Vertical position does not matter.

Figure 2. The equalization curve and control slider learned for "bright" in a single session.

Once a participant selects a descriptive term and a sound file, they are asked to "rate how 'your word' that sound is" using the interface shown in 1. The participant then rates 40 examples of the chosen audio file, each modified with an equalization curve. Fifteen curves are repeats of prior examples, to test for rating consistency. Figure 1 shows the interface for rating examples.

From these rated examples, the system learns the relative boost/cut to apply each of 40 frequency bands. These bands have equivalent rectangular bandwidth (ERB) [3] derived center frequencies. We use the method from [15] to learn the EQ curve from user responses. This method treats each frequency band as independent and correlates changes in user ratings with changes in gain for that band.

Positive correlation indicates a boost, and negative correlation a cut. The result is a 40-band EQ curve for the descriptor learned from the participant. The system uses the learned equalization curve to build a slider that lets the participant manipulate the sound in terms of the descriptor (Figure 2).

After teaching SocialEQ an audio descriptor and trying the control slider, participants were asked to complete a question survey that assessed their background, listening environment, and experience using SocialEQ.

4. DATA COLLECTION

For a data collection of this size, an on-site data collection was not feasible. We instead recruited participants through Amazon's Mechanical Turk. We had 633 participants who participated in a total of 1102 training sessions (one session per learned word). We paid participants $1.00 (USD) per session, with the possibility of up to a $0.50 bonus, determined by the consistency of their equalized audio example ratings. While the quality of loudspeakers could not be controlled, over 92% of the participants reported listening over either headphones or large speakers (rather than small/laptop speakers).

4.1 Inclusion Criteria for Sessions

Before analyzing the results, we first removed sessions by participants who seemed to not put effort into the task. The mean time to teach the system the definition of a single descriptor was 292 seconds (SD=237). We removed all sessions where the participant completed the task in less than 60 seconds. We also removed all sessions where the participant gave the default rating for more than 5 out of the 40 examples. We also removed any session where the participant responded "no" to the survey question: "Was the listening environment quiet?".

Recall that 15 of the 40 examples were repeats in any session. This let us test for consistency of user responses to audio examples when teaching SocialEQ. We measured consistency using Pearson correlation between the ratings of the test and repeated examples. The median consistency across sessions was 0.41 (95% CI [0.39, 0.44]).

Only sessions with consistency above 0 were retained. This left 481 participants who taught the machine in 731 sessions (Individual participants were allowed to teach more than one descriptor to the system. The maximum number of descriptors a participant taught was 9.).

5. RESULTS

Since we have data on hundreds of words taught by hundreds of participants, we are not able to describe it all in detail here. We have made the data available for use by the research community at `http://socialeq.org/data`.

5.1 Definitions

descriptor: An adjective (e.g. *warm*) taught to the system in at least one session.

session: A participant teaches SocialEQ a descriptor, tries the learned controller, and completes the survey.

relative spectral curve (RSC): A set of relative gains (i.e. boosts or cuts) on the 40 ERB frequency bands. To make RSC comparable, every RSC is normalized by putting gain values in standard deviations from the mean value of the 40 bands.

user-concept: The RSC learned in a single session (e.g. the session where Bob teaches 'warm' to SocialEQ).

descriptor definition The set of user-concepts that all share a descriptor. A definition may be vague or precise depending on how much agreement there is between user-concepts that share the descriptor. Figure 3 shows *deep* and *sharp*.

5.2 The descriptors

In the 731 sessions, there were 324 unique descriptors. The descriptors taught most frequently are listed in Table 1. Of the 324 words, 91 occurred two or more times. The most popular descriptor was *warm*, but note that there was a bias for participants to teach the system *warm* due to its use as the example in the instructions (see Section 3).

Table 1. The 10 most common descriptors contributed

Rank	Descriptor	Sessions
1	warm	57
2	cold	25
3	soft	24
4	loud	22
5	happy	19
6	bright	16
7	harsh	15
8	soothing	14
9	heavy	11
10	cool	11

5.3 Representing equalization concepts

In this paper, we make the assumption that participants judged each equalization example relative to the unprocessed source rather than judging the absolute spectrum of each equalization example. We therefore have chosen to represent equalization concepts in terms of relative changes in each frequency band rather than the resulting spectrum after equalization. This allows us to compare equalization concepts from varying source material played on varying loudspeakers. In Figure 3, we show the distribution of RSCs collected for two example descriptors: *deep* and *sharp*. Each column shows the distribution of learned values for the corresponding ERB-spaced frequency band.

Note that except for one outlier participant, there was fairly high agreement for the meaning of *sharp*. This is interesting, since most musicians are taught that *sharp* relates to relative pitch, rather than the spectral characteristics of a sound. Our data indicate *sharp* also has other connotations that relate to timbre.

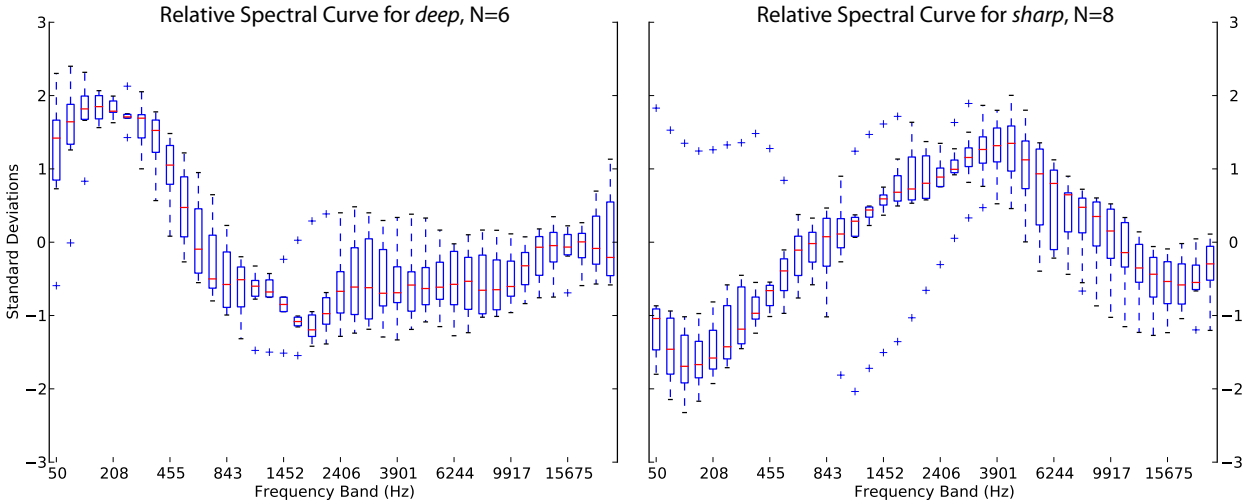

Figure 3. Per frequency band boxplots of RSCs for the descriptors *deep* and *sharp*, learned in 6 and 8 sessions respectively. In each ERB-spaced frequency band, the center line is the median, the box represents 50% of the data, the whiskers represent the remaining 50%. The pluses represent outliers.

5.4 Actionable equalization descriptors

Table 2. Top 10 descriptors taught to the system by at least 4 people, ranked by *mean slider rating*, which ranges from -3 (Strongly Disagree) to 3 (Strongly Agree)

Rank	Word	Mean Response
1	relaxing	2.75
2	quiet	2.60
3	hot	2.50
4	hard	2.50
5	heavy	2.36
6	smooth	2.33
7	deep	2.33
8	bright	2.31
9	soothing	2.31
10	mellow	2.29

As stated in Question 1 in Section 1, one one goal of this paper is to determine what audio descriptors describe goals achievable by an audio equalizer (i.e. the audio descriptor is an equalization descriptor). One way to answer this is to simply look at the *mean slider rating*: the mean response to the survey statement, "The final (control) slider captured my target audio concept." Participants were asked to respond on a 7-level Likert scale coded from -3 (Strongly Disagree) to 3 (Strongly Agree). Table 2 shows the 10 descriptors with the highest mean response to this question that were contributed by at least 4 participants.

The learning approach used by SocialEQ [15] has an inherent bias toward learn smooth equalization curves and has difficulty learning curves with narrow boosts or cuts or frequency relationships that are non-linear or dependent. Therefore, *mean slider rating* is a sufficient but not necessary condition to determine whether the descriptor is actionable by an equalizer.

5.5 Agreed upon equalization descriptors

Table 3. Top 10 descriptors ranked by the *agreement score* described in Section 5.5

Rank	Word	Agreement Score
1	tinny	0.294
2	pleasing	0.222
3	low	0.219
4	dry	0.210
5	metallic	0.195
6	quiet	0.188
7	deep	0.164
8	hollow	0.160
9	light	0.131
10	warm	0.130

The second question we would like to answer is "How widely-agreed-upon is the meaning of an EQ descriptor?". For some words, the meaning of a descriptor, as embodied in the RSC learned in a particular session, may vary significantly from person to person. We want to find which descriptors vary the least from person to person, or rather which descriptors have the most widely-agreed-upon meanings. To answer the question we looked at the total variance (i.e. the trace of the covariance matrix) of RSCs within a descriptor definition. This can be written simply as the sum of the variance in each RSC frequency-band for the user-concepts in a descriptor definition:

$$\text{trace}(\Sigma)_{descriptor} = \frac{1}{N}\sum_{k=0}^{39}\sum_{n=0}^{N-1}(x_{n,k} - \mu_k)^2 \quad (1)$$

where N is the number of user-concepts in the descriptor definition, k is the index of the frequency band, x is the RSC for user-concept n, and μ_k is the mean of frequency

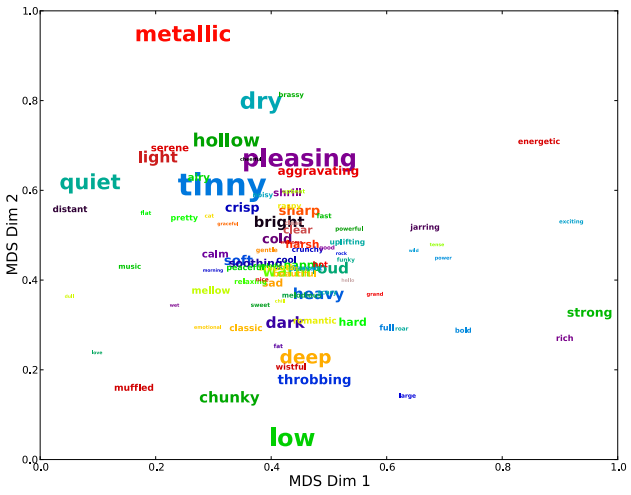

Figure 4. 2-dimensional multi-dimensional scaling (MDS) plot of the descriptors. Their font size positively correlates to their agreement score from Equation 2.

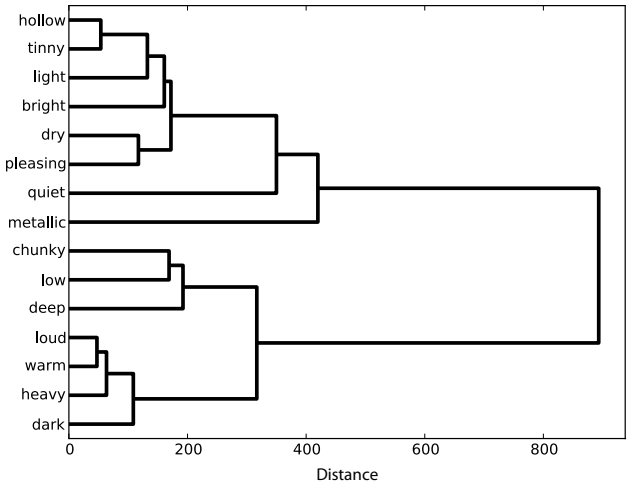

Figure 5. Hierarchical clustering of the 15 descriptors with the highest agreement scores.

band k over the N user-concepts in the descriptor definition.

If we then divide the natural logarithm of the number of user-concepts ($\log(N)$) by this value, i.e.:

$$agreementscore = \frac{log(N)}{\text{trace}(\Sigma)_{descriptor}} \quad (2)$$

we have an *agreement score* that takes into account both total variance and the popularity of the descriptor. We used $\log(N)$ to linearize the number of user-concepts since the frequency with which a descriptor was taught the system was distributed similarly to Zipf's law [10]. When we rank the descriptors by this score, we discover which descriptors have more agreement amongst the participants. The top ten descriptors ranked by this score are shown in Table 3.

5.6 Audio descriptor synonyms

To answer the last question, "What EQ descriptors are true audio synonyms", we compared the descriptor definitions using a distance function.

Table 4. Synonyms of high agreement descriptors (see Table 3) found through comparing descriptor definitions

descriptor	synonyms
light	tinny, crisp
tinny	hollow, crisp, light, shrill, bright, cold, raspy
deep	throbbing, dark
hollow	tinny, dry, shrill, pleasing

To compare learned descriptor definitions, we wanted a distance measure that would: 1) allow for varying number of user-concepts per descriptor definition; 2) allow for multi-modal distributions; and 3) take into account the uncertainty of the descriptor definition. Therefore, we modelled each descriptor definition as a probability distribution over the user-concepts for the descriptor, and we then compared descriptor definitions using an approximation of the symmetric KL-divergence.

The steps to calculate the distance between two descriptor definitions are:

1. Model each user-concept as a Gaussian distribution, $\mathcal{N}(\mu_i, \Sigma_i)$, where μ_i is the RSC of the user-concept and Σ is a diagonal covariance matrix in which the variance for each frequency-band is set by $\sigma^2_{i,k} = (\sigma_k - \sigma_k r_i)^2$ where σ_k is the sample standard deviation of frequency-band k for RSCs of *all* descriptors, and r_i is the ratings consistency for the session that learned user-concept i. Here we are using the ratings consistency as a measure of the uncertainty of the user-concept, mapping a consistency range of $[0, 1]$ to a per-frequency-band variance range of $[\sigma_k, 0]$.

2. Then model each descriptor definition as follows:

$$P(x) = \frac{1}{N} \sum_{i=0}^{N-1} \mathcal{N}(\mu_i, \Sigma_i) \quad (3)$$

   where $\mathcal{N}(\mu_i, \Sigma_i)$ is the distribution for the i^{th} user-concept.

3. We then use a symmetric Monte Carlo approximate KL-divergence [7] to compare two descriptor definition models.

Using this distance measure, we computed the distance between every pair of descriptor definitions. With these distances, we can map and visualize the relationships of descriptor definitions. In Figure 4, we placed the descriptor definitions in a two-dimensional space by using metric multi-dimensional scaling [1], each descriptor definition is scaled by its agreement score so that well-defined descriptors are larger. To get a better sense of the relationships of the 15 high agreement descriptor definitions listed in Table 3, we performed agglomerative hierarchical clustering using the "group average" algorithm [5] and plotted the dendrogram in Figure 5. From this you can see two clusters

formed with descriptors one might typically associates as opposites: *bright/dark*, *quiet/loud*, *light/heavy*. From this we can also see relationships of descriptors which may not have been obvious such as *pleasing* being closely associated with *dry*.

In Table 4, we list a few high agreement descriptors along with their synonyms through comparing descriptor definitions. We considered two words synonyms if their distance was within the first percentile of all the pairwise distances. Also, only descriptors definitions that consisted of at least two user-concepts and had at least a 1.0 *mean slider rating* (as described in Section 5.4) were included.

6. CONCLUSION

In this work, we analyzed the equalization descriptors that were taught to a web-based equalization learning system, SocialEQ. We developed methods to determine which descriptors map well to equalization, which have broadly-agreed upon meaning, and which are synonymous. During analysis we found both expected and unexpected relationships and definitions of descriptors, but more importantly we developed the ground work of an intelligent audio production system that responds to the descriptive language of the user. With a large collection of equalization descriptors and the techniques described in this paper, we have a map of equalization descriptors which a system could use to determine whether a goal is achievable with an equalizer and whether it needs to learn an individual's equalization concept or can use an agreed upon concept.

This work was supported by grant by National Science Foundation Grant No. IIS-1116384.

7. REFERENCES

[1] I. Borg and P. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Verlag, 2005.

[2] A. Disley and D. Howard. Spectral correlates of timbral semantics relating to the pipe organ. *Speech, Music and Hearing*, 46:25–39, 2004.

[3] B. Glasberg and B. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(12):103–138, 1990.

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

[5] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer New York, 2001.

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

[7] J. Hershey and P. Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *Proc. ICASSP, 2007*.

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

[9] E. Lukasik. Towards timbre-driven semantic retrieval of violins. In *Proc. of International Conference on Intelligent Systems Design and Applications, 2005*.

[10] C. Manning, P. Raghavan, and H. Schtze. *Introduction to information retrieval*. Cambridge University Press Cambridge, 2008.

[11] S. McAdams, S. Winsberg, S. Donnadieu, G. Soete, and J. Krimphoff. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58(3):177–192, 1995.

[12] S. Mecklenburg and J. Loviscach. subjeqt: controlling an equalizer through subjective terms. In *Proc. of CHI '06 Extended Abstracts on Human Factors in Computing Systems, 2006*.

[13] G. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[14] D. Reed. Capturing perceptual expertise: a sound equalization expert system. *Knowledge-Based Systems*, 14(12):111–118, 2001.

[15] A.T. Sabin, Z. Rafii, and B. Pardo. Weighting-function-based rapid mapping of descriptors to audio processing parameters. *Journal of the AES*, 59(6):419–430, 2011.

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

[17] J. Scott, R. Migneco, B. Morton, C. Hahn, P. Diefenbach, and Y. Kim. An audio processing library for mir application development in flash. In *Proc. ISMIR, 2010*.

[18] R. Shepard. Geometrical approximations to the structure of musical pitch. *Psychological Review*, 89(4):305–333, 1982.

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

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

[21] E. Toulson. A need for universal definitions of audio terminologies and improved knowledge transfer to the audio consumer. In *Proc. of The Art of Record Production Conference, 2003*.

[22] A. Zacharakis, K. Pastiadis, G. Papadelis, and J. Reiss. An investigation of musical timbre: Uncovering salient semantic descriptions and perceptual dimensions. In *Proc. of the ISMIR, 2011*.